SYSTEMATIC ACCELERATION OF RADICAL DISCOVERY AND INNOVATION IN SCIENCE AND TECHNOLOGY

BY

DR. RONALD N. KOSTOFF
OFFICE OF NAVAL RESEARCH
800 N. QUINCY ST.
ARLINGTON, VA 22217
Phone: 703-696-4198
Fax: 703-696-3098
Internet: kostofr@onr.navy.mil

DISCLAIMER

(The views in this paper are solely those of the author, and do not necessarily represent the views of the Department of the Navy, or any of its components)

KEYWORDS

Discovery; Innovation; Science and Technology; Text Mining; Literature-Based Discovery; Literature-Assisted Discovery; Information Retrieval; Unconnected Disciplines; Disparate Disciplines; Interdisciplinary; Multidisciplinary; Solicitations; Special Issues; Workshops; Roadmaps; Advisory Panels; Review Panels

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Systematic Acceleration of Radical Diiscovery and Innovation in Science and Technology	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research Dr. Ronald N. Kostoff 800 N. Quincy Street Arlington, VA 22217**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**83**	

ABSTRACT

Literature-based discovery (LBD) is a systematic two-component approach to bridging unconnected disciplines (front-end component, back-end component) based on text mining procedures. LBD allows potentially ***radical*** discovery and innovation (***radical*** discovery and innovation is used in the sense of discovery and innovation arising from unexpected insights originating in very disparate disciplines) to be hypothesized. Classically, the LBD front-end component has been used to identify the pool of potential discovery and innovation candidates, and the LBD back-end component has been used to hypothesize the potential discovery and innovation based on literature analysis alone (1).

In this report, a systematic two-component approach to bridging unconnected disciplines and accelerating potentially radical discovery and innovation (based wholly or partially on text mining procedures) is presented. The front-end component has similar objectives to those in the classical LBD approach, although it is different mechanistically and operationally. The front-end component in the present report will systematically identify technical disciplines/ technologies (and their associated leading experts) that are directly or indirectly-related to solving technical problems of high interest. The back-end component in the present report is actually a family of back-end techniques, only one of which shares the strictly literature-based analysis of the classical LBD approach. These multiple back-end techniques will identify potential radical discovery and innovation for many different applications. In the present report, the two-component techniques that use strictly literature-based analysis for the back-end are termed ***literature-based discovery***. The two-component techniques that do not focus on literature-based analysis for the back-end are termed ***literature-assisted discovery***.

In the front-end, those very directly related disciplines/ technologies normally associated with the technical problems of interest can be viewed as 'internal' disciplines, and their associated literature can be viewed as the core literature. Those disciplines/ technologies related less directly (or indirectly) to the technical problems of interest can be viewed as 'external' disciplines, and their associated literature can be viewed as the expanded literature. These 'external' disciplines (represented by literature and people) serve as potential sources of radical discovery and innovation, and are examined in the back-end.

Specifically, in the ***literature-assisted discovery*** operational mode, these 'external' discipline experts could be used as:

1. Recipients of solicitation announcements (BAA, SBIR, MURI, journal Special Issue calls for papers, etc),
2. Participants in Workshops, Advisory Panels, Review Panels, and Roadmaps,
3. Points of Contact for Field Science Advisors, Foreign Field Offices, Program Officer site visits, and potential transitions

In the above ***literature-assisted discovery*** roles, the 'external' discipline experts would extrapolate the insights and principles from the 'external' disciplines to solve problems of interest related to the core literature. In some applications, use of these 'external' experts will not require structural changes to many organizations' present business operations. In other applications, proper use of these 'external' experts will have characteristics of 'disruptive technologies', mainly due to the requirement to properly handle the infusion of large numbers of concepts and insights from very disparate disciplines. The complete ***literature-based discovery*** approaches could be used to complement the ***literature-assisted discovery*** (people-based) approaches. Potential advantages of using these literature-assisted and literature-based approaches include more radically innovative science and technology (S&T), improved global leveraging of S&T, improved coordination with domestic S&T sponsoring agencies, and technical journals acting more proactively to stimulate radical discovery and innovation. Additionally, these literature-based or literature-assisted approaches could offer S&T investors better insight into the potential of cutting-edge technologies.

DEFINITIONS

Discovery is ascertaining something previously unknown or unrecognized. Innovation reflects the metamorphosis from present practice to some new, hopefully "better" practice. It can be based on existing non-implemented knowledge, discovery of previously unknown information, discovery and synthesis of publicly available knowledge whose independent segments have never been combined, and/ or invention. In turn, the invention could derive from logical exploitation of a knowledge base, and/ or from spontaneous creativity (e.g., Edisonian discoveries from trial and error). (2).

INTRODUCTION

Discovery and innovation are the cornerstones of frontier research. One of the methods for generating radical discovery and innovation in a target discipline is to use principles and insights from very disparate disciplines (to the target discipline) to solve problems in the target discipline. Unfortunately, identifying these linkages between the disparate and target disciplines, and making the subsequent extrapolations has tended to be a very serendipitous process. Until now, there has been no fully systematic approach to bridging these unconnected target and disparate disciplines. The present report describes a systematic approach, or more specifically, variants of a systematic approach (based wholly or partially on text mining procedures) for making these connections. One of the virtues of these specific approaches is that most of them can easily be integrated into the operational processes of science and technology (S&T) sponsoring organizations, or research performing organizations. However, some of these approaches do have characteristics of 'disruptive technologies', due to the additional effort required to properly integrate large numbers of concepts representing many disparate disciplines.

There are many examples where enhancement/ acceleration of discovery/ innovation requires insights and knowledge from 'external' technical disciplines, sometimes very disparate disciplines. One could envision a solution to 'mine detection' that exploits the remote detection of markers of nitrogen homeostasis in the presence of clinical disorders. In this case, the 'internal' technical discipline is that ordinarily associated with 'mine detection', while the 'external' technical disciplines would be those associated with specific aspects of remote (or possibly in-situ) detection not normally associated with 'mine detection'. The real challenge is to have a systematic process that identifies these 'external' disciplines starting from the 'internal' disciplines (thereby retaining some indirect thread of connectivity between the 'internal' and 'external' disciplines), and then extrapolates the insights and knowledge from these 'external' disciplines to solve problems in the 'internal' discipline or technology of interest.

The challenge has become more critical due to increasing specialization and effective isolation of technical/ medical researchers and developers (4). As research funding and numbers of researchers have increased substantially over the past few decades, the technical literature has increased substantially as a result. Researchers/ developers struggle to keep pace with their own

disciplines, much less to develop awareness of other disciplines. Thus, we have the paradox that the *expansion of research* has led to the *balkanization of research*! The resulting balkanization serves as a barrier to cross-discipline knowledge transfers, and retards the progress of discovery and innovation (4).

There are two main text mining avenues for extrapolating knowledge and insights from one discipline/ technology to another: ***literature-based discovery*** and ***literature-assisted discovery***. The ***literature-based discovery*** approach uses technical experts to access and examine the literature from 'external' disciplines to help solve problems in the 'internal' discipline. The ***literature-assisted discovery*** approach uses technical experts from 'external' disciplines in a variety of interactive and/ or independent creative modes for the same purpose.

The main thesis of this report is that the scientific community has not made adequate use of these 'external' discipline sources of knowledge to accelerate potentially radical discovery and innovation. Further, very substantive quality enhancements to funding agency S&T programs, individual research projects, journal Special Issues, and multi-disciplinary teams and organizations are possible at relatively small marginal costs, if we can systematically improve access to the limitless sources of 'external' discipline/ technology information.

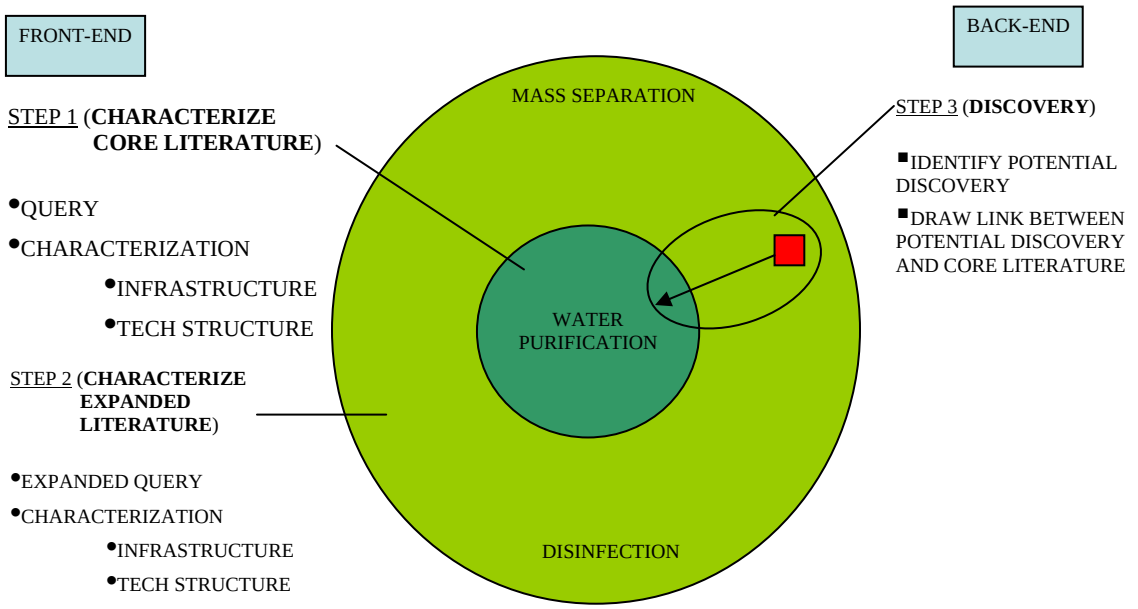
RADICAL DISCOVERY AND INNOVATION CONCEPT

Text mining is the extraction of useful information from large volumes of text (3). The author's text mining effort over the past decade has been developing methods to systematically access external sources of information that could contribute to problem solving for specific technical disciplines, technologies, systems, operations, or technical problems in general (see Appendix References, Kostoff, 1991a-2005h)). These methods have been integrated to form the following systematic approach for accelerating radical discovery and innovation.

Figure 1 contains a schematic of our text mining approach to discovery. The inner circle represents the core literature of the problem to be solved. In the example for Figure 1, the problem to be solved is identifying 'improved' alternatives to existing water purification technologies, where 'improved' could encompass any combination of lower cost, lower energy use, lower

maintenance, higher reliability, lighter weight, and improved modularity for field assembly. Thus, the core literature is the existing water purification literature. The annular region between the inner and outer circles represents literatures directly or indirectly related to the core literature.

Figure 1



The discovery process presented in this report is divided into two components, a front-end and a back-end.

FRONT-END

The front-end component contains two major steps: characterization of the core literature, and characterization of the expanded literature, including technical experts associated with this literature. In step 1, a query to retrieve the core literature is developed iteratively (the Appendix describes our approach to query development for both steps 1 and 2 in some detail). Once the core literature has been retrieved with this query, it is subject to text mining. Bibliometrics provides the technical infrastructure (key authors/ institutions/ countries/ journals, etc) of the core literature, and computational linguistics provides the technical structure (technical thrusts, hierarchical taxonomies) of the core literature. Step 1 can be viewed as a characterization of the core literature, and reflects the scope of many of our mono-technology text mining studies to date.

In step 2, the query developed in step 1 is generalized and expanded, again iteratively (see the last pre-Summary/ Conclusions section in the Appendix [section IV-K] for some practical techniques to generate the expanded query). This expanded query will retrieve records from literatures directly and indirectly related to the core literature. Insights and principles from these disparate literatures/ technical disciplines can be extrapolated to solve problems of the core literature. Thus, in the example on Figure 1, the water purification literature is expanded to cover all of mass separation and disinfection. Insights from very disparate mass separation and disinfection approaches can be extrapolated to solve problems in water purification.

BACK-END

The back-end component contains the discovery step, which itself contains two sub-components. The first sub-component is identification of potential discovery and innovation candidates from the expanded literature, and the second sub-component is drawing the linkages between the potential discovery/ innovation candidates and the core literature. There are many ways to identify potential discovery and innovation candidates, and to draw the subsequent linkages. These techniques differ mainly by the approach mechanics and the types of people used to identify the discovery and innovation candidates. The two main discovery and innovation approach types (Literature-Based, Literature-Assisted) are described now.

A. Literature-Based Discovery

We can use systematic techniques to identify potential discovery and innovation based strictly on the 'external' literature, an approach known as literature-based discovery (LBD-1, 2). LBD is useful in the planning and concept identification phases of the S&T development cycle. The literature-based approach can be viewed as a very sophisticated type of literature survey, and represents a somewhat different way of doing business for most S&T sponsoring agencies, researchers, and technical journals.

B. Literature-Assisted Discovery

We can identify technical experts associated with these 'external' disciplines, and then have them focus their expertise on solving problems of interest from the 'internal' disciplines. This literature-assisted people-based approach could easily be incorporated into most S&T sponsoring agencies' existing operational procedures. However, in some applications, proper handling of the infusion of large numbers of concepts and insights from

disparate disciplines will acquire the characteristics of 'disruptive technologies'.

Thus, the differences between paths A and B are in the 'back-end', in 1) how the linkages between the 'external' and 'internal' disciplines are made, and 2) who makes the linkages.

The ultimate goal should be incorporation of both approaches in parallel, to exploit the strengths of each approach while eliminating the weaknesses. This synergy would provide the *comprehensiveness and objectivity* of the completely literature-based approach coupled with the *interaction and feedback* of the literature-assisted people-based approach.

With respect to the literature-assisted approach, how would the 'external' discipline experts be incorporated into different components of the overall research enterprise's operations, for the purpose of enhancing and accelerating discovery and innovation? The following *options* are a sample of what is possible.

1) Solicitations – Science and Technology Sponsoring Organizations
Government agencies and private foundations generate numerous solicitations for proposals and/ or new ideas for solving problems. In the United States, these include Broad Agency Announcements (BAAs), Small Business Innovation Research (SBIR) solicitations, and other similar types. For the Federal agencies, these solicitations are usually advertised in some widely available forum (e.g., FedBizOpps) and, in parallel, some announcements of the solicitation are disseminated to technical experts deemed knowledgeable about the topic. ***These targeted announcements are extremely important, since they insure that the recipient will be aware of the solicitation***. The numbers of announcements are usually modest, because of limited prior access to potentially relevant communities.

OPTION 1. The *first option* is that the 'external' discipline experts identified through the text mining, as well as the comprehensive list of 'internal' discipline experts, constitute the bulk of the announcement distribution list. In this way, their expertise in a directly or indirectly-related discipline could be brought to bear on solving the sponsor's problem of interest, with their motivation amplified by the potential for funding, if successful. These 'internal' and 'external' discipline experts would also serve as the gateway to identifying additional technical experts (in these

disciplines) not associated with the particular literatures accessed (e.g., through common professional societies, institution sub-divisions, attendance at conferences with the experts identified through the strictly literature approach), and thereby adding these additional technical experts to the problem-solving process. Use of the expanded notification list could result in an order of magnitude more proposals, and perhaps two orders of magnitude more potentially radical discovery and innovation proposals.

Potential consequences ('side-effects') resulting from this new approach to solicitations include 1) **substantial** increases in numbers of proposals (***as we have demonstrated successfully***), 2) need to expand diversity of reviewers' technical disciplines to insure interdisciplinary proposals receive balanced evaluation (4), and 3) need to facilitate/ stimulate discovery process by improved notification instructions. Consequences 1) and 2) could have modest 'disruptive technology' characteristics, especially if very large numbers of proposals from experts in many disparate disciplines are received.

2) Solicitations – Science and Technology Journals

Many technical specialty journals are structured on centuries-old research archival and dissemination models. Their scope is 'stove-piped' about quite narrow themes. This parochialism is further compounded by the increasing influence of Impact Factor as a publication metric target, restricting the types of articles published to increasingly narrower bands. The publication trend is toward more narrowly discipline-focused articles that will receive high citations. The trend is away from articles that encompass very diverse disciplines, may not receive high citations on average due to their inter-disciplinary nature (4), but could stimulate more radical discovery and innovation.

My recent citation studies of specific technical journals and of specific multi-journal technical disciplines show graphically the narrow dispersion of highly cited paper types, especially in the technical specialty journals. This trend needs to be reversed if the technical journals are to assume their rightful positions as engines of innovation. The following paragraphs offer one approach for reversing this trend.

Most technical journals produce Special Issues periodically. These Special Issues tend to focus on a single topic, and usually have recognized experts on the specific topic present their perspectives. The papers tend to focus on

comprehensiveness of coverage about the specific topic, rather then venturing into very disparate disciplines in a search for discovery.

There are two main avenues by which Special Issue papers are solicited. One is the Guest Editor (usually a recognized expert in the Special Issue topic) inviting other recognized experts known to him/ her. The second is the Guest Editor/ journal placing 'call for papers' ads in prior issues of the journal, or other closely-related topic-centric journals. In both cases, the result is the same: papers centered closely about the topic of interest.

However, the Special Issue concept could be expanded to emphasize radical discovery and innovation. As in the science and technology sponsoring organization example, notification of the projected Special Issue could be sent to the technical experts identified in the front end of the text mining discovery study. These experts would be encouraged to submit papers for the Special Issue that involved extrapolation of insights and principles from their own technical specialties to solving problems in the Special Issue topic. In this operational mode, the technical journals would serve proactively as the engines for radical discovery and innovation.

OPTION 2. The *second option* is that the 'external' discipline experts identified through the text mining, as well as the comprehensive list of 'internal' discipline experts, constitute the bulk of the journal Special Issue announcement distribution list. In this way, their expertise in a directly or indirectly-related discipline could be brought to bear on solving the Special Issue's particular problem of interest, with their motivation amplified by the potential for journal publication, if successful.

The potential consequences from this new mode of Special Issue operation mirror those of the first option: more papers than normal, greater topical diversity in papers, and need to facilitate discovery and innovation. It is strongly recommended that the Special Issue be expanded in size from the normal journal practice, to accommodate the expected increase in number of submittals. Sponsorship of such Special Issues by the science and technology funding organizations seems appropriate. Modest honoraria could also be provided to authors as further motivation for participation.

3) Advisory Panels

Government agencies and private organizations convene numerous advisory panels or groups of independent advisors for the purpose of providing expert

technical advice on problems of present and future interest. Unfortunately, many of these advisory groups are somewhat parochial, both in terms of technical scope and people, thereby limiting the breadth of their recommendations.

OPTION 3. The *third option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a significant portion of the members of these groups.

4) Workshops

Government agencies and private organizations conduct numerous workshops for the purpose of generating new project ideas and directions. Many of these workshop participants are somewhat parochial, both in terms of technical scope and people. Additionally, venture capital organizations, and other components of Wall Street, have great needs for workshops that could provide insight on the potential of emerging technologies. It would be useful for these organizations to know whether the core technology of interest is amenable to improvement, and whether some of the indirectly-related technologies can offer potential solutions across many different core technologies.

OPTION 4. The *fourth option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a significant portion of the participants at these workshops (1). If the workshops are conducted in tandem with the solicitation processes above, then the results from the solicitations could be used to narrow the pool of candidates for these workshops. The workshop attendees could be drawn from the solicitation announcement recipient group that submitted proposals to the sponsoring organization solicitations, or the solicitation announcement recipient group that submitted papers to the technical journal solicitations. At a minimum, the members of these groups had sufficient insight to perceive how concepts from their areas of expertise could be extrapolated to solve problems of interest in the core technology.

5) Review Panels

Government agencies and private organizations conduct numerous review panels during the execution of their S&T programs. Many of these review panels are somewhat parochial, both in terms of technical scope and people. This limits discussion of the breadth of approaches that could be used to achieve the program objectives.

OPTION 5. The *fifth option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a portion of the reviewers.

6) Roadmaps

Some government agency and private organization programs generate technology roadmaps (see (6) for a description of technology roadmaps) as part of their planning processes, and/ or as part of their review processes. Many of these roadmap development teams have limited perspectives, both in terms of technical scope and people. The breadth of these roadmaps is limited by the breadth of their developers.

OPTION 6. The *sixth option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a significant portion of the roadmap development team. As in the workshop option, the solicitation step could be used to filter the candidate pool for prospective roadmap development team members.

7) Points of Contact

Many government agency components require points of contact (POCs) for obtaining information to solve problems. These include the Field Science Advisors for military organizations, Foreign Field Offices of government agencies, Program Officers for site visits, and Program Officers to identify potential transitions. In practice, many of these POCs accessed derive from limited personal knowledge, both in terms of technical scope and people. The breadth of the information obtained from the POCs is limited by their breadth of expertise.

OPTION 7. The *seventh option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a very significant portion of the POCs accessed in practice. Again, as in the workshop option, the solicitation step could be used to filter the candidate pool for prospective POCs.

8) Organization and Team Structuring

Technical teams and organizations can be structured to maximize the potential for radical discovery and innovation, based on the principles presented in the paragraphs above. Multi-disciplinary groups/ structures such as Integrated Product Teams (IPTs), Multi-Disciplinary Research

Programs of the URI (MURIs), and Cooperative Research and Development Agreements (CRADAs) could be assembled based on the technical disciplines and technical experts identified at the front-end of the discovery process above. Organizations such as Centers of Excellence with a defined core competency, and large laboratories with multiple core competencies, could be structured to incorporate the technical disciplines surrounding their core identified at the front-end of the discovery process.

OPTION 8. The *eighth option* is that the 'external' (and 'internal') disciplines identified through the text mining constitute a significant portion of the teams and organizations. Where possible, the solicitation step could be used to filter the candidate pool for prospective team and organization members.

SUMMARY AND CONCLUSIONS

I have proposed the identification and exploitation of diverse literatures, and their representative experts, to help solve problems of interest through potentially radical discovery and innovation. The approach is based on our demonstrated text mining techniques. Their essential element is development of comprehensive and precise queries for retrieving the expanded literature of potential discovery candidates (see Appendix for query development techniques), followed by exploitation of these retrieved literatures and their associated technical expert representatives.

I have identified a number of pathways by which these literatures and people could be integrated with present business practices, or could be integrated with slight modifications if desired. ***This group of 'external' literature accession techniques has the highest benefit/ cost ratio of any techniques I know for enhancing and accelerating radical discovery and innovation.***

Additionally, the "structural holes" research of Professor Burt (6), which identifies 'structural holes' as the weakly-linked or non-linked region between different technical disciplines, has shown that most innovations studied have been the result of drawing on insights from unconnected, sometimes very disparate, disciplines/ technologies. These findings support the thesis that is the basis for my proposal above. Translated into practice, the following important guidelines can be drawn.

1. If we are interested in meeting short-term deadlines efficiently with specific well-defined technology products, then homogeneous well-coordinated and long-standing groups are most useful.
2. If we are interested in radical discovery and innovation, including innovation in the advanced technology demonstration of system integration sense, then we need to incorporate people *we don't know* representing disciplines with which *we are not familiar*. It is difficult to get ***'out-of-the-box'*** thinking from people who have spent their careers ***'in-the-box'***!

The generic literature-based discovery approach provides a systematic and objective guide to selecting the most appropriate disciplines and people for accelerating potentially radical discovery and innovation.

REFERENCES – MAIN TEXT

1. Swanson, DR, Smalheiser, NR. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*. 91(2).
2. Kostoff, RN. Stimulating innovation. *International Handbook of Innovation*. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 2003.
3. Hearst, M. Untangling text data mining. In the Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.
4. Kostoff, RN. Overcoming specialization. *BioScience*. 52:10. 937-941. 2002.
5. Kostoff, RN, and Schaller, RR. Science and technology roadmaps. IEEE Transactions on Engineering Management. 48:2. 132-143. May 2001.
6. Burt, RS. Structural holes: The social structure of competition. Harvard University Press. Cambridge, MA. 1992.

APPENDIX – DEVELOPMENT OF QUERIES FOR INFORMATION RETRIEVAL

I. Abstract

This Appendix describes recent advances to an iterative full-text information retrieval approach based on relevance feedback with term co-occurrence and query expansion. The method generates search terms from the language and context of the text authors, and is sufficiently flexible to apply to a variety of databases. It provides improvement to the search strategy and related results as the search progresses, adding relevant records to the information retrieved and subtracting non-relevant records as well. Finally, it allows maximum retrieval of relevant records with high signal-to-noise ratio by tracking marginal utility of candidate query modification terms using a semi-automated optimization procedure.

Simulated Nucleation, the name given to this relevance feedback variant, is derived in concept from the growth of materials. In Simulated Nucleation for information retrieval, a small core group of documents relevant to the topic of interest is retrieved using a test query. Unique characteristics of these core documents are identified from bibliometric (authors, journals, institutions, sponsors, citations) and computational linguistics (phrase frequency and phrase proximity) analyses. Patterns of bibliometrics and phrases and phrase combinations in existing fields are identified, the test query is modified by human experts with new search term combinations that follow the newly identified patterns, and the process is repeated. In addition, patterns of bibliometrics and phrases and phrase combinations that reflect extraneous non-relevant material are identified, and search terms that have the ability to remove non-relevant documents from the retrieved database are added to the modified query. Thus, Simulated Nucleation operates in a self-correcting cybernetic homeostatic mode with the assistance of expert analysis, and continually expands the coverage and improves the quality of the retrieved core database. This iterative procedure continues until convergence is obtained, where relatively few new documents are found or few non-relevant documents are identified, even though new search terms are added. An application is described of developing, from the Science Citation Index (SCI), a database of journal articles focused on the technical domain of textual data mining (TDM).

II. Introduction

II-A. Magnitude of Technical Data Explosion

Global expenditures for science and technology (S&T) range from $500B/year to $1T/year, depending on one's definition of S&T (Kostoff, 2003a). Federal agencies and industrial organizations with broad mission areas have eclectic S&T needs. Results from many different types of S&T are required in order to accomplish the overall objectives of the sponsoring organizations. In addition, because of the interlocking nature of S&T, results from many different types of S&T efforts are required to produce advances in any specific area.

However, any particular Federal agency or industrial organization can afford to sponsor only a very small fraction of the S&T necessary to provide the technical foundations for accomplishing its broader mission objectives. This means that it is imperative for any Federal agency or industrial organization (that requires significant technological advances to accomplish its broad mission objectives) to maintain awareness of the S&T being performed globally. This continual awareness will allow the agency or company to leverage and exploit the results of externally-sponsored S&T for its own, and national, benefit.

In practice, the S&T programs of these sponsoring organizations are managed by a wide variety of technical managers. To effectively manage their programs, exploit and leverage external programs, and remain at the cutting edge of S&T, the managers need to:

1) understand what S&T was done in the past, to both exploit it presently and not repeat mistakes that were made in the previous development;
2) understand what S&T is being conducted presently, to both leverage existing programs for optimal resource use and avoid duplication;
3) understand what S&T is planned to be conducted, to allow a) strategic budgetary planning for future S&T transitions; b) planning strategic cost-sharing in areas of common interest; and c) withdrawal of planned budgets from areas of peripheral interest that will be addressed elsewhere.

The S&T managers require this information not only in S&T areas directly related to their technologies of interest, but in allied and disparate technical fields as well. These supporting technical areas can serve as sources of innovation and discovery for advancing the prime technical areas, and can help remove the underlying

critical path barriers that serve as roadblocks to progress along the primary technical paths.

The managers need to understand this information at two levels of resolution:

1) Macro level – initially, they need to understand the information in its larger global context, to understand the temporal and spatial trends, and the vertical and horizontal relationships. At this level, the longer-range strategic concerns can be addressed.
2) Micro level – once the overall structural relationships are understood, the managers need to understand a more limited information subset at the detailed level. This is the resolution necessary for daily operations, and allows the shorter-range tactical concerns to be addressed.

The managers need access to a variety of sources for this information, to gain the full spectrum of perspectives on available S&T. These sources include human contacts, literature, multi-media, and physical sources. The present document addresses an improved information retrieval approach to accessing one of these sources, the literature. While the specific applications of the method have focused on a relatively narrow range of semi-structured text databases (e.g., the Science Citation Index of basic research papers, the Engineering Compendex of applied research and technology papers, the Medline database of medical research papers, the NTIS Technical Reports of government-sponsored S&T), the method is sufficiently flexible to address a much wider group of text data sources.

In particular, the method could be used to extract information from literatures characteristic of the temporal segments described above. Some of these literatures include:

1) Past S&T – Archived primary and secondary literatures, reports, memos, Web documents.
2) Present S&T – Recently published papers and reports, internal reports, program and project narratives, open and internal files, Web documents.
3) Future S&T - Published and internal plans, published and internal roadmaps, program descriptive narratives, S&T funding proposals submitted electronically.

The information retrieval method could be used as the cornerstone of a process that would both extract information directly from the text sources as well as use the preliminary extracted information as a gateway to the other data sources. For

example, simple processing of the very comprehensive information retrieved by the present method will identify S&T performers, journals, organizations, and sponsors. These sources can then be contacted to provide a more personal type of information retrieval, and supplement the literature-based approach extensively.

II-B. Growth of Available Data

Over the past decade, with the growth and expansion of electronic storage media, there has been a virtual explosion of multi-media data readily available. In particular, use of CD-ROMs and the Internet has provided overwhelming data resources to the user community.

The Web version of the Science Citation Index (SCI) accesses over 32 million technical documents from 5600 technical journals, and presents this information in semi-structured textual format. In 2004, the SCI added approximately 1.1 million new technical documents. In order to convert disorganized data of this magnitude to structured useful information, a variety of computer-assisted decision aids have been developed. To extract useful information from large volumes of semi-structured and unstructured textual material, sophisticated TDM techniques have been generated. But what particular types of problems can TDM address?

II-C. Textual Data Mining for Extracting Information from Text

Generically, TDM can support selection of prolific and related experts for workshops and review panels, and prolific organizations and performers as site visit candidates. It can provide the underlying data required for roadmap development, strategic planning, and international policy assessment. TDM can automatically identify novel information groupings. The expert analysts supported by the TDM algorithms can identify new technical insights, promising R&D opportunities, and cross-database linkages (requirements -→ opportunities, diseases -→ cures, etc).

Specific issues that arise repeatedly in the conduct of R&D management (the motivation for the present work), and that could potentially be addressed by TDM, include:

- What R&D is being done globally;
- Who is doing it;
- What is the level of effort;
- Where is it being done;

- What are the major thrust areas;
- What are the relationships among major thrust areas;
- What are the relationships between major thrust areas and supporting areas, including the performing and archiving infrastructure;
- What is not being done;
- What are the promising directions for new research;
- What are the innovations and discoveries?

These issues can be divided into two categories: infrastructure (who, where) and technical (what). What are some of the conceptual techniques used to address each category?

II-D. Generic Components of Textual Data Mining

To address these questions, TDM techniques typically have four major generic components, under the broad definition of TDM used operationally by the author. There is an *information retrieval* component to select the appropriate raw textual data on which the information processing will be performed. There is a *bibliometrics* component that identifies the people, archival, institutional, and regional infrastructure of the topical domain being analyzed. There is a *computational linguistics* component that extracts topical themes of interest, and relationships among these themes and the infrastructure components. Finally, there are visualization and/ or other types of *information display* components that summarize the TDM analyses and results for the users/ customers. The latter component is mechanistic relative to the more intrinsically fundamental first three components.

While all four data mining components are important for a high quality useful product, the comprehensiveness and focus of the information retrieval component are central and fundamental to the quality of the results and the latter three components. All the sophisticated bibliometrics and computational linguistics processing cannot compensate for insufficient or unfocused base data. Unfortunately, standard information retrieval approaches tend to cast either too wide a net or too fine a net to make optimal use of the available data. As a result, the potential user is either overwhelmed with extraneous data, or is uninformed about existing valuable information.

II-E. Problems with Present Information Retrieval Approaches

The key problem with most standard search approaches is that the analyst is required to hypothesize the search terms in the context of the application, rather than use the database to provide the search terms appropriate to the context in which they are actually imbedded. For those databases or search engines that provide an on-line dictionary of terms, the analyst is still unable to ascertain the context in which those phrases are employed, and therefore cannot predict whether the theme of the article targeted by the search term used is the theme desired. The search approaches that try to approximate context better by weighting different search terms, and then using a figure of merit to select documents that effectively contain more and a wider variety of the weighted search terms, still have the limitations of being based on analyst hypotheses.

For general language databases, automated text retrieval using linguistic rules and supporting on-line dictionaries could provide marginally acceptable results for some classes of users. For highly technical S&T databases, the focus of the text retrieval techniques discussed in the present Appendix, automated text retrieval results in poor retrieval performance. Highly specialized and comprehensive dictionaries are required for good performance, since the technical jargon appears as a foreign language to the processing algorithms. Natural language processors have severe limitations when applied to highly specialized technical terms. Sense ambiguity of many technical terms further compounds the problem of selecting terms based on context. As confirmed by the present study, very few text mining studies have been reported for analysis of specialty technical disciplines.

II-F. Requirements for High Quality Information Retrieval Approaches

A high quality information retrieval approach should not only be able to yield the search terms from the language and context of the authors rather than from the language of the searcher, but should have other desirable properties. It should be able to work efficiently on a variety of different databases, and on different fields or combinations of fields in semi-structured databases. Some databases have keywords; most don't. Some databases provide authors; some don't. Some databases provide references; some don't. Analogous to a neural network's operation, a high quality information retrieval approach should be able to improve the search strategy and results as time proceeds. It should be able to produce a higher signal (comprehensiveness of documents retrieved) with the passage of time, but also a higher signal-to-noise ratio (focus of documents retrieved) as well. Finally, a high quality approach should allow more records in allied S&T fields to be retrieved, and allow relevant records in disparate fields to be retrieved. The latter capability has high potential for generating innovation and discovery from

disparate disciplines (Kostoff, 1999b). This Appendix describes an information retrieval approach that has these desirable properties, and more.

II-G. Definition of Quality in Information Retrieval Context

Finally, the term ‘quality’ must be used in its larger context, in the degree to which the information retrieval supports the larger study objectives. The following example illustrates the problem in defining information retrieval quality, and shows that more than the simplistic precision and recall metrics typically used to gauge quality in the Text Retrieval Conference (TREC) studies are required to provide a pragmatic operational definition of quality.

The parent article to the present Appendix (Kostoff, 1997a) focused on the use of computational linguistics imbedded in an iterative relevance feedback procedure. The final product is a query that will retrieve documents with two aggregate characteristics; the maximum number of relevant documents will be retrieved, and the ratio of relevant to non-relevant documents will be very large; i.e., the signal-to-noise ratio will be large.

Quality in the context of information retrieval requires three conditions. Two of these conditions are the aggregate characteristics mentioned above. The third condition derives from the definition of 'relevant', and requires the desired definition of 'relevant' to be incorporated into the query development process. The operational meaning of 'relevant' depends on the objectives of the query. Is the purpose of the query to retrieve all the papers in:

*a target technical field (Step 1 in Figure 1), and/ or
*allied technical fields as well, and/ or
*very disparate technical fields (Step 2 in Figure 1) that have the potential to provide innovative new insights to advance the target technical field (Kostoff, 1999c).

Each of these purposes defines a very different concept of 'relevant', and would result in very different numbers of 'relevant' documents being retrieved.

Typical R&D literature surveys have none of these three quality conditions. Most queries consist of a few key words fairly closely associated with the desired narrow target literature, with minimal (if any) iterative steps (Kostoff, 2001a). The results will either contain substantial noise if the search terms are relatively broad, or will be very limited if the search terms are narrowly focused. Some iterative

approaches will provide substantial numbers of records with high signal-to-noise ratio using a constrained definition of relevant; i.e., not accessing the disparate literatures from which innovative ideas could potentially flow. Rarely, if ever, are all three necessary conditions for a high quality information retrieval fulfilled. Why is this?

Probably the main reason is time and cost. Information retrieval/ text mining efforts (e.g., Kostoff, 1999c, 1999d) have shown that an iterative process that incorporates a broad scope of 'relevant' disciplines to the target discipline requires the participation of

1) a technical domain expert(s) and
2) either a computational linguistics expert(s) or a documented procedure using computational linguistics tools.

There is substantial judgement and interpretation required by at least one expert at each iterative step, and this effort directly translates into significant resource expenditures. The downside of not expending sufficient resources to obtain a high quality product is that allied and related literatures that could serve as the engines of innovation are not accessed.

As an example of the level of effort required for a reasonable quality query, the author, in conjunction with two technical domain experts, developed a query related to the hydrodynamic flow over solid bodies (for examining flow around ships). The objective of the study was to comprehensively characterize the core literature (Step 1), not to expand the query for discovery (Step 2). Three iterative steps were required; each step required the technical expert(s) to read many hundreds of the retrieved records in order to identify those that were relevant and non-relevant . Then, computational linguistics analyses (Kostoff, 1997a) were performed on both the relevant and non-relevant records to identify phrase patterns and relationships characteristic of the relevant records and the non-relevant records. Substantial time and judgement were required to select the appropriate phrases unique to the relevant records and the non-relevant records, and then modify the query accordingly using the key phrases identified. Approximately 200 terms were contained in the final query. Even then, the process could have continued for more iterations, but it was not considered cost-effective given the time and resource constraints of the specific study.

In a more recent example, query development was done for the topic of water purification, for both core literature characterization and discovery objectives. The

query to retrieve the core literature was on the order of 50-100 terms (based on the definition of core literature), and the query to retrieve the expanded literature was on the order of 500-1000 terms (depending on the database accessed). As in the case above, more iterations could have been used, but they did not appear cost-effective from a marginal utility perspective.

III. Background

As part of a larger project on TDM, the author performed a survey of the information retrieval literature to identify efficient query concepts (Kostoff, 2000d, 2001a). Most of the information retrieval papers tended to be very abstract, contained little detail that would provide guidance in conducting an actual query, and appeared marginally related to the much less esoteric methods used by actual information retrieval practitioners (librarians, etc.). In addition, most of these published information retrieval papers focused on the use of search terms derived from extrinsic sources, rather from the language used by the database text authors. These deficiencies would probably limit the efficiency and comprehensiveness of the search process and final results.

Since the present author has been developing database analysis techniques based on Term Co-occurrences over the past few years, it was decided to see whether these techniques could be adapted and expanded to address the literature survey problem. To achieve the objective of developing the high quality information retrieval approach whose properties were described in the previous section, the Term Co-occurrence retrieval component already developed by the author would have to be embedded in a Relevance Feedback structure with Query Expansion. An ancillary objective was to develop and articulate this novel information retrieval approach for use by real-world practitioners.

The remainder of this Background section outlines the major efforts that have been performed in Term Co-Occurrence, Co-Word Analysis, Query Expansion, and Relevance Feedback, and addresses some of their limitations. These four topics are not distinct, but have substantial overlap, and this is reflected in the historical survey. Then, the computational linguistics technique developed by the author, Database Tomography, is outlined, and the initial application of its information retrieval variant, Simulated Nucleation, is summarized.

III-A. Term Co-Occurrence and Co-Word Analysis

Term Co-occurrence has its roots in co-word analysis and computational linguistics. Co-word analysis utilizes the proximity of words and their frequency of co-occurrence in some domain (sentence, paragraph, paper, etc.) to estimate the strength of their relationship. When applied to the literature in a technical field, co-word analysis allows a map of the relationship among technical themes to be constructed. A history of co-word analysis applied to research policy issues, its origins in computational linguistics, and its limitations due to previous dependence on the sole use of key words and index words, can be found in Kostoff (1993a).

III-B. Term Co-Occurrence and Query Expansion

Term Co-occurrence in information retrieval can be used to expand on an initial query, and the additional query terms allow the retrieval of relevant documents that would not have been retrieved with the initial query. These additional terms could also be used to remove irrelevant documents. Traditionally, this form of Query Expansion has been carried out by means of thesauri and controlled vocabularies. The construction of these is time-consuming and expensive. Additionally, these extra terms are analyst-generated, rather than generated from the phrases used in the searched database.

Studies related to the use of Term Co-occurrence in information retrieval can be traced back to at least the 1960s (Maron and Kuhns, 1960; Stiles; 1961; Lesk, 1969). These early experiments demonstrated the potential of Term Co-occurrence data for the identification of search term variants. They eventually led to the conclusion that Query Expansion provided the greatest improvement in performance when the original query gave reasonable retrieval results, whereas expansion was less effective when the original query had performed badly. This is in accord with the Association Hypothesis: "If an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is also likely to be good at this" (Van Rijsbergen, 1979).

The use of Term Co-occurrence in information retrieval was motivated initially by the heuristic observations that searchers and authors tend to use different terms to describe the same information, and consequently a number of related query terms are required for increased search effectiveness and efficiency. These observations were confirmed later by experiments (Furnas et al, 1987; Gomez et al, 1990).

Additional work on Query Expansion related to Term Co-occurrence has been based on probabilistic models of the retrieval process and has tried to relax some of the strong assumptions of term statistical independence that normally need to be

invoked if probabilistic retrieval models are to be used. Results have been mixed; it was not found possible to obtain consistent improvements in performance by the use of any of the Query Expansion methods (Croft and Harper, 1979; Robertson and Sparck Jones, 1976; Smeaton and Van Rijsbergen, 1983) . This has not been the experience of the present author, and may be due to the sole reliance by the authors on natural language expressions from the database for Query Expansion terms rather than utilization of external sources for these additional terms.

Early 1990s work in the use of Term Co-occurrence for Query Expansion exploited available computing power to generate thesauri automatically (Crouch, 1990; Rasmussen, 1992). A later series of reported studies has used automatic indexing and co-occurrence analysis, performed on parallel computers, to generate a domain-specific thesaurus automatically [Chen et al, 1997, 1998]. A recent paper examined the factors affecting the performance of global query expansion based on term co-occurrence data, and suggested ways to maximize the retrieval effectiveness (Chung and Lee, 2004).

III-C. Relevance Feedback and Query Expansion

Relevance Feedback is a controlled process for query reformulation. The main idea consists of choosing important terms, or expressions, attached to certain previously retrieved documents that have been identified as relevant by the users, and of enhancing the importance of these terms in a new query formulation (Salton and Buckley, 1990). Analogously, terms included in previously retrieved non-relevant documents could be de-emphasized in any future query formulation. A comprehensive overview of the pre-1992 literature on Relevance Feedback is contained in Harman (1992).

Initially, the Relevance Feedback implementations were designed for queries and documents in vector form; i.e., query statements consisting of sets of possibly weighted search terms used without Boolean operators (Rocchio, 1971; Salton, 1971) . Since the 1980s, Relevance Feedback methods have been applied also to Boolean query formulations, where the process incorporates term conjuncts (derived from previously retrieved relevant documents) into revised query formulations (e.g., Salton et al, 1985)

In the mid-1990s, Relevance Feedback approaches with probabilistic information retrieval based on document components have been incorporated into artificial neural networks (e.g., Kwok, 1995). This approach recognizes the intrinsic Relevance Feedback operation of artificial neural networks, and the natural

application to the information retrieval process. Performance with feedback improved substantially over the no feedback case.

Another important series of mid-1990s studies, that has an important bearing on the present Appendix, focused on determining the retrieval effectiveness of search terms identified by users and intermediaries from retrieved items during term Relevance Feedback. Results show that terms selected from particular database fields of retrieved items during term relevance feedback (TRF) were more effective than search terms from the intermediary, database thesauri or users' domain knowledge during the interaction. The study concludes that more focus on the practice of database searching and on the origins of the terms used for the feedback process is necessary (Spink, 1995, 1996, 1998).

Also in the mid-1990s, use of local context analysis (Xu and Croft, 1996), that combines global analysis (Jing and Croft, 1994; Callan and Croft, 1995) and local feedback (Attar and Fraenkel, 1977), has generated effective information retrieval results. In this combined approach, noun groups are used as concepts and concepts are selected based on co-occurrence with query terms. Concepts are chosen from the top-ranked documents, similar to local feedback, but the best passages are used instead of whole documents. An algorithm is used to rank the concepts, and the query is then expanded.

A recent paper surveys relevance feedback techniques, examining both automatic techniques, in which the system modifies the user's query, and interactive techniques, in which the user has control over query modification. It also considers specific interfaces to relevance feedback systems and characteristics of searchers that can affect the use and success of relevance feedback systems (Ruthven and Lalmas, 2003).

In summary, the overwhelming majority of the reported Relevance Feedback studies for Query Expansion appear to have focused on the mathematical operations and cognitive aspects of the feedback process. While this focus has resulted in many of the process innovations and advances, it has been limited in eliminating the inconsistency of the Relevance Feedback process results. Relatively few innovative approaches have been applied to identifying more appropriate sources of expansion terms. The present Appendix focuses mainly on the expansion term sources.

III-D. Database Tomography (Kostoff, 1995b)

Classical co-word analysis applied to index/ key words for the purpose of science and technology (S&T) evaluation does not allow the richness of the semantic relationships in full text to be exploited, and it is restricted to formally published papers (see Kostoff, 1993a, for a more detailed history of co-word analysis, and its evolution to a decision aid for research evaluation). In order to allow any form of free text to be used, Database Tomography (DT) was developed.

In 1990-1991, experiments were performed at the Office of Naval Research that showed the frequency with which phrases appeared in full text narrative technical documents was related to the main themes of the text (Kostoff, 1991). The phrases with the highest frequencies of appearance represented the main, 'pervasive' themes of the text. In addition, the experiments showed that the physical proximity of the phrases was related to the thematic proximity. These experiments formed the basis of DT.

The DT method in its entirety requires generically three distinct steps. The first step is identification of the main themes of the text being analyzed. The second step is determination of the quantitative and qualitative relationships among the main themes and their secondary themes. The final step is tracking the evolution of these themes and their relationships through time. The first two steps will be summarized now. Time evolution of themes has not yet been studied.

First, the frequencies of appearance in the total text of all single word phrases (e.g., Matrix), adjacent double word phrases (e.g., Metal Matrix), and adjacent triple word phrases (e.g., Metal Matrix Composites) are computed. The highest frequency technical content phrases are selected by topical experts as the pervasive themes of the full database.

Second, for each theme phrase, the frequencies of phrases within some domain centered about the theme phrase are computed for every occurrence of the theme phrase in the full text, and a phrase frequency dictionary is constructed. Past DT studies have used a domain whose length was a fixed number of words from the theme phrase (usually +- 50 words. More recent on-going studies are examining the use of domains with syntactic boundaries, such as article Abstracts, paragraphs, and sentences.

This phrase frequency dictionary contains the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses are performed by the topical expert for each dictionary (hereafter called cluster) yielding, among many

results, those sub-themes closely related to and supportive of the main cluster theme.

Third, threshold values are assigned to the numerical indices, and these indices are used to filter out the phrases most closely related to the cluster theme. However, because numbers are limited in their ability to portray the conceptual relationships among themes and sub-themes, the qualitative analyses of the extracted data by the topical experts have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full text analysis allow an understanding of the theme inter-relationships not heretofore possible with previous text abstraction techniques (using index words, key words, etc.).

At this point, a variety of different analyses can be performed. For databases of non-journal technical articles (Kostoff, 1991b, 1993b, 1993c, 1994b, 1994c) the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting sub-thrust areas (both high and low frequency) to the high-frequency themes. For the more recent studies in which the databases are journal article abstracts and associated bibliometric information (authors, journals, addresses, etc), the final results have also included relationships among the technical themes and authors, journals, institutions, etc (Kostoff et al, 1997b, 1997c, 1998a, 1999a, 2000a, 2000c, 2000g, 2001b, 2001g, 2002a, 2003c, 2003d, 2003l, 2003n, 2003p, 2003s, 2004a, 2004c, 2004d, 2004f, 2004h, 2004i, 2004j, 2004l, 2004m, 2004n, 2004o, 2005a, 2005b, 2005f, 2005h).

Especially in the information retrieval phase, the balance of effort of the more recent studies has shifted from computer-centric (where the primary emphasis was on the computer results, and the secondary emphasis was on the expert analysis of the computer results) to expert-centric (where the primary emphasis is on expert analysis of the computer results and raw data, and the computer results serve to augment the capabilities of the expert). Expert-centric S&T text mining provides an in-depth understanding/ identification of the technical concepts and their inter-relationships, whereas the computer-centric approach focused on the more superficial level of context-free phrases.

III-E. Utilization of Database Tomography for Information Retrieval

Database Tomography applied to information retrieval has all the desirable qualities of the high quality information retrieval process listed in the Introduction, and more. It will operate on any textual database in any language. Its minimal

requirement is text only, but will be enhanced with use of titles, keywords, references, bibliometrics, etc. It works directly from the language of the authors of the database's contents, and improves the search strategy and product with time. Its value increases as the size of the desired retrieval increases.

Simulated Nucleation (Kostoff et al, 1997a), the name given to the form of Database Tomography adapted to information retrieval, derives in concept from the growth of materials. A core nucleus is developed, the properties of this nucleus are identified, and then similar material is added onto the nucleus as time develops until the desired amount of material is obtained. Impurities can be removed at any time during the growth process, but most impurities are removed at the end of the growth process. The growth process is then terminated.

In Simulated Nucleation for information retrieval, the purpose is to provide a tailored database of retrieved documents, that contains all relevant documents from the larger literature. The final product may also contain a minimal amount of non-relevant documents. In the initial step of Simulated Nucleation, a small core group of documents mainly relevant to the topic of interest is identified by the topical domain experts, analogous to the core nucleus of material described above. An inherent assumption is then made that the bibliometric and phrase patterns and phrase combinations characteristic of this relevant core group would be found to occur in other relevant documents. Therefore, these bibliometric and phrase patterns and phrase combinations can be used to expand the search query, again analogous to the addition of similar material to the core nucleus of material described above.

While both bibliometrics and computational linguistics have been used in Simulated Nucleation to identify unique characteristics of each category, the bulk of the development effort has concentrated on the computational linguistics. Therefore, the bulk of the remainder of this paper will address the computational linguistics.

There are two major Simulated Nucleation approaches for expanding the number of relevant documents and contracting the number of non-relevant documents. The first is a manually intensive approach, requiring the reading of many sample Abstracts to separate the relevant from non-relevant documents, and then identify candidate query terms from computational linguistics analysis of each document category. The second is a semi-automated approach, using computer-based clustering techniques for separating the relevant from non-relevant records (e.g., see Hearst (1996) and Zamir (1999) for examples of clustering approaches to

separate relevant from non-relevant documents), but still requiring manual identification of candidate query terms from computational linguistics analysis of each separate document category. The initial reported effort was on the first approach (Kostoff et al, 1997a); recent efforts have focused on the second approach (Kostoff et al, 2005a).

In the first approach, the main algorithmic components of Database Tomography, phrase frequency and phrase proximity analyses, operate on this core group of documents. Patterns of phrase combinations in existing fields are identified, new search term combinations that follow the newly identified patterns are generated, and the process is repeated. In addition, patterns of phrase combinations that reflect extraneous non-relevant material that may have been introduced are identified, and search terms that have the ability to remove non-relevant documents from the database are inserted. Thus, Simulated Nucleation operates in a self-correcting cybernetic homeostatic mode, and continually expands the coverage and improves the quality of the core database. This iterative procedure continues until convergence or low marginal utility is obtained (Kostoff et al, 2004a), where relatively few new documents or non-relevant documents are found even though new search terms are added.

The information retrieval approach described in this paper is based on Term Co-occurrence with Relevance Feedback for Query Expansion. It employs the DT algorithms, and can be used by a wide variety of analysts with little training to provide highly efficient retrieval.

In the remainder of this Background section, the adaptation and utilization of Database Tomography for information retrieval is presented. In the next section, some very recent additions to Simulated Nucleation that identify high impact query modification terms more efficiently are described in detail.

III-F. Application of Simulated Nucleation for Information Retrieval

In 1997, a paper was published (Kostoff, 1997a) describing Simulated Nucleation, its intellectual heritage, and its application to a data mining study focused on utilization of near-earth space. The interested reader is strongly urged to read this document, in order to obtain a detailed understanding of the operation of Simulated Nucleation. Since that time, Simulated Nucleation has been used in a number of published data mining studies in different topical areas (Kostoff et al, 1997b, 1997c, 1998a, 1999a, 2000a, 2000c, 2000g, 2001b, 2001g, 2002a, 2003c, 2003d, 2003l, 2003n, 2003p, 2003s, 2004a, 2004c, 2004d, 2004f, 2004h, 2004i, 2004j,

2004l, 2004m, 2004n, 2004o, 2005a, 2005b, 2005f, 2005h), and in shorter non-published efforts as well. Much deeper insights about the operation and benefits of Simulated Nucleation have been obtained, and a number of algorithmic upgrades have been made to enhance the power and efficiency of Simulated Nucleation. Some of these insights, and the algorithmic upgrades, will be described in the following sections, in the context of a Simulated Nucleation operational protocol.

IV. Recent Upgrades to Simulated Nucleation

Most of the work on Simulated Nucleation described above was performed in the 1995-1997 time frame. Since that time, there has been much experience gained from working with Simulated Nucleation, and a much greater appreciation of its importance in a TDM study. In particular, the role that the technical domain expert plays in constructing a query is crucial, and the need for queries of substantial size is understood better.

Furthermore, additional algorithms have been generated to increase the power of Simulated Nucleation and to make it more efficient as well. The updated process that is used presently, and the operation of these new algorithms within that process, will be described in some detail. The main example used is a TDM study of the discipline of TDM.

IV-A. Overview of Updated Process

The operational objective of Simulated Nucleation is to generate a query that will have the following characteristics:

*Retrieve the maximum number of records in the technical discipline of interest
*Retrieve substantial numbers of records in closely allied disciplines
*Retrieve substantial numbers of records in disparate disciplines that have some connection to the technical discipline of interest
*Retrieve records in aggregate with high signal-to-noise ratio (number of desirable records large compared to number of undesirable records)
*Retrieve records with high marginal utility (each additional query term will retrieve large ratio of desirable to undesirable records)
*Minimize query size to conform to limit requirements of search engine(s) used

To achieve these objectives, the Simulated Nucleation process has been improved and updated, and now contains the following steps:

*Definition of study scope
*Generation of query development strategy
*Generation of test query
*Retrieve records from database; select sample
*Divide sample records into relevant and non-relevant categories
*Perform computational linguistics on each category
*Use new algorithms to identify phrases unique to each category
*Use new algorithms to identify phrase combinations unique to each category
*Use new algorithms to identify marginal value of adding each phrase and phrase combination to query
*Construct modified query
*Repeat process until convergence obtained

Each of these steps will now be described in more detail, and the upgrades and improvements will be emphasized.

IV-B. Definition of Study Scope

The process begins with a definition of the scope of the study by all participants. Within the context of this scope, an initial query is constructed. (Since each iterative step follows the same procedure, only one iterative step from the TDM study of TDM will be described.) Queries are scope dependent. Typically, if a new scope is defined, a new query must be developed. However, due to the iterative nature of Simulated Nucleation, if the scope becomes more focused within the overall topical domain as the study proceeds, the new scope can be accommodated within succeeding iterations. Such a scope sharpening did occur during the course of the illustrative TDM example, and the accommodation of the new scope within the iterative process will be summarized now.

For the TDM study example, the initial TDM scope was defined as retrieving records related to textual data mining in the larger context; i.e., including information retrieval. In the physical ore mining analogy, this is equivalent to mining and processing the ore. As the study proceeded, the scope was restricted to documents that focused on understanding and enhancing the quality of the TDM process, as opposed to using standard TDM approaches to perform specific studies.

IV-C. Generation of Query Development Strategy

The next step in the upgraded Simulated Nucleation process is generation of a query development strategy. Past experience with Simulated Nucleation has shown that the structure and complexity of a query are highly dependent on:

*the objectives of the study;
*the query construction philosophy;
*the contents of the database selected;
*the fields queried within the database;
*the background and perspectives of the expert(s) constructing the query;
*the parametric tradeoffs made (comprehensiveness of records, signal-to-noise ratio, number of iterations, resources available)

These query dependencies are taken into account when structuring the initial query. Different initial queries will eventually evolve to similar final queries through the iterative process. However, higher quality initial queries will result in a more streamlined and efficient iterative process.

IV-C-i. Relation of Query Structure to Study Objectives, Construction Philosophy, and Database Selected

Specifically, one of the key findings from a FY98 ONR TDM pilot program was that, in general, a separate query had to be developed for each database examined (Kostoff and Geisler, 1999c; Kostoff et al, 2000c). Each database accesses a particular culture, with its unique language and unique types of documentation and expression. A query that optimizes (retrieves large numbers of desirable records with high signal-to-noise ratio (relevance)) for one database within the context of the study objectives may be very inadequate for another database.

For example, one of the FY98 pilot program studies focused on the R&D of the aircraft platform (Kostoff et al, 2000c). The query philosophy was to start with the generic term AIRCRAFT, then add terms that would expand the numbers of aircraft R&D records (mainly journal paper Abstracts) retrieved and would eliminate records not relevant to aircraft R&D. Two databases were queried, the Science Citation Index (SCI-a database accessing basic research records) and the Engineering Compendex (EC-a database accessing applied research and technology records). The SCI query required 207 terms and three iterations for an acceptable signal-to-noise ratio, while the EC query required 13 terms and one iteration to produce an even better signal-to-noise ratio. Because of the technology focus of the EC, most of the records retrieved using an aircraft or helicopter type query term focused on the R&D of the platform itself, and were aligned with the

study goals. Because of the research focus of the SCI, many of the records retrieved focused on the science that could be performed from the aircraft platform, rather than the R&D of the platform, and were not aligned with the study goals.

Therefore, no adjustments were required to the EC query, whereas many negation terms (NOT Boolean terms) were required for the SCI query to eliminate aircraft records not aligned with the main study objectives. It is analogous to the selection of a mathematical coordinate system for solving a physical problem. If the grid lines are well aligned with the physical problem to be solved, the equations will be relatively simple. If the grid lines are not well aligned, the equations will contain a large number of terms required to translate between the geometry of the physical problem and the geometry of the coordinate system.

IV-C-ii. Relation of Query Structure to Database Fields Selected

In TDM, queries, as well as follow-on computational linguistics analyses, could provide misleading results if applied to one database field only. The text fields (Keywords, Titles, Abstracts) are used by their originators for different purposes, and contain different levels of information content and detail (Kostoff et al, 2000c, 2000g, 2004f, 2004n, 2004o, 2005f). The query and other computational linguistics results could provide a different picture of the overall discipline studied based on which field was examined.

As an example, in the aircraft study referenced previously, queries were applied to all text fields (Keywords, Titles, Abstracts) simultaneously. However, follow-on phrase frequency analyses for TDM were performed on multiple database fields to gain different perspectives. A high frequency Keyword focal area concentrated on the mature technology issues of longevity and maintenance; this view of the aircraft literature was not evident from the high frequency Abstract phrases. The lower frequency Abstract phrases had to be accessed to identify thrusts in this mature technology/ longevity/ maintenance area. Queries applied to these different fields separately would have resulted in a similar bias in records retrieved.

Keywords are author/ indexer summary judgements of the main focus of a technical paper's contents, and represent a higher level description of the contents than the actual words in the paper/ Abstract. Thus, one explanation for the difference between the conclusions from the high frequency Keywords and Abstract phrases of the aircraft study is that the body of non-aircraft maintenance Abstract phrases, when considered in aggregate from a gestalt viewpoint, are viewed by the author/ indexer as aircraft maintenance/ longevity oriented.

However, while there may be a difference in high frequency phrases between the two data sources, there may be far less of a difference when both high and low frequency phrases are considered. Thus, a second possible explanation is that, in some technical areas in different databases, there is more diversity in descriptive language employed. Rather than a few high frequency phrases to describe the area, many diverse low frequency phrases are used. This could result from the research encompassing a wider spectrum of smaller effort topics.

It could also result from the absence of a recognized discipline, with its accepted associated language. This would reflect the arbitrary combination of a number of diverse fields to produce the technical area, with the associated numerous but low frequency thrusts. Another explanation for the difference between the Keywords and Abstracts perspectives of aircraft maintenance is that maintenance and longevity issues are politically popular now, and the authors/ indexers may be applying (consciously or subconsciously) this 'spin' to attract more reader interest.

Also, the Abstract phrases from the aircraft study contained heavy emphasis on laboratory and flight test phenomena, whereas there was a noticeable absence of any test facilities and testing phenomena in the Keywords. Again, the indexers may view much of the testing as a means to the larger end. Their keywords reflect their perception of the ultimate objectives or applications, rather than the detailed approaches for reaching these objectives as would be derived from the Abstracts or even from the full paper's text. However, there was also emphasis on high performance in the Abstract phrases, a category conspicuously absent from the Keywords. In fact, the presence of mature technology and longevity descriptors in the Keywords, coupled with the absence of high performance descriptors, provided a very different picture of aircraft literature research from the presence of high performance descriptors in the Abstract phrases, coupled with the absence of mature technology and longevity/maintenance descriptors. Queries applied to either field alone would retrieve records that reflected this bias.

IV-C-iii. Relation of Query Structure to Expert(s) Perspectives

The TDM analytical procedure in which Simulated Nucleation is imbedded, and the query construction, are not independent of the analyst's domain knowledge; they are, in fact, expert-centric. The computer techniques play a strong supporting role, but they are subservient to the expert, and not vice versa. The computer-derived results help guide and structure the expert's analytical processes; the computer output provides a framework upon which the expert can construct a

comprehensive story. The final query and study conclusions, however, will reflect the biases and limitations of the expert(s). Thus, a fully credible query and overall analysis requires not only domain knowledge by the analyst(s), but probably domain knowledge representing diverse backgrounds (i.e., multiple experts). It was also found useful in each FY98 TDM pilot program study to incorporate a generalist with substantial experience in constructing queries and analyzing different technical domains. This person could identify efficient query terms and unique patterns for that technical domain not evident to the more narrowly focused domain experts.

IV-C-iv. Summary of Query Development Difficulties

The bottom line on query development that resulted from the FY98 pilot program is that constructing an R&D database query that will retrieve sufficient technical documents to be of operational use is not a simple procedure. It requires:

*close interaction with technical experts;
*in-depth understanding of the contents and structure of the potential databases to be queried by the study performers;
*sufficient technical breadth of the study performers in aggregate to understand the potentially different meanings and contexts that specific technical phrases could have when used in different technical areas and by different technical cultures (e.g., SPACE SATELLITES, SATELLITE CLINICS, SATELLITE TUMORS);
*an understanding of the relation of these database contents to the problem of interest; and
*substantial time and effort on the part of the technical expert(s) and supporting information technologist(s).

These stringent and complex requirements run counter to the unfounded assertions being promulgated presently by the algorithm developers and vendors in the information technology community: sophisticated tools exist that will allow low-salaried non-experts to perform comprehensive and useful data retrieval and analysis with minimal expenditures of time and resources.

While query construction is complex for semi-structured textual databases such as the aforementioned SCI and EC, it is far more complex for an unstructured data source such as the World Wide Web. The Web is a conglomeration of many types of data, with no central structure to the records, with data of widely varying contents and quality and verification, and unknown completeness and coverage. Based on the experiences from the FY98 TDM pilot program, there is no evidence

that a rigorous query of high quality and utility (comparable to those developed using Simulated Nucleation and the semi-structured textual SCI and EC databases) could be made of the highly unstructured Web as it exists now and in the foreseeable future.

IV-C-v. Types of Query Construction Philosophy

There are two generic types of query construction philosophy that have been used with Simulated Nucleation.

IV-C-v-1. Generic Term Initialization

One philosophy starts with relatively broad terms, and builds the query iteratively. Many of the additional terms, however, may be non-relevant to the scope of the study due to the multiple meanings the more general terms may be assigned. Some query modification procedure is required to eliminate non-relevant records. For example, in the aircraft R&D study (Kostoff et al, 2000c), this general approach was used. The query started with AIRCRAFT, and then was modified to remove terms that would result in retrieving aircraft records not related to the R&D of the aircraft platform. While the emphasis of these later iterations was reduction of non-relevant records, there were terms added to the query that would retrieve new records.

IV-C-v-2. Specific Term Initialization

The other philosophy starts with relatively specific terms, and builds the query iteratively as well. Most of the additional query terms will retrieve relevant records. Because of the specificity of the query terms, records relating to the more general theme and scope of the study may be overlooked.

IV-C-v-3. Iteration Temporal Sequence Considerations

Also, within both philosophies, if multiple iterations are used, the focus is different for each iterative step in the temporal sequence. The earlier iterations emphasize adding query terms to expand the number of relevant records retrieved, while the later iterations emphasize modifying the query to reduce the number of non-relevant records retrieved.

IV-C-v-4. Impact of Resource Limitations on Query Construction Philosophy

Each iteration allows new related literatures to be accessed, and additional relevant records to be retrieved. However, additional time and money are required for each added iteration, because of the intense analysis required per iteration.

In practice, the two main limiting parameters to the length of a study are number of iterations and resources available. Two practical cases of interest will be addressed.

IV-C-v-4-a. Minimize Iterations

The first case results from severe resource constraints. In this case, the objective is to minimize the number of iterations required to develop the query subject to a threshold signal-to-noise ratio on retrieved records. The strategy for a single iteration query is to generate a test query (initial guess), categorize the retrieved records into relevant and non-relevant bins, apply computational linguistics to each bin, and select only those phrases and phrase combinations that are strongly characteristic of the relevant bin for the modified query. The ratio for phrase selection cutoff will be determined by the marginal utility of each phrase as a query term. The resulting records retrieved with this modified query will have very high signal-to-noise ratio, as confirmed by sampling a few records retrieved with this modified query. However, their coverage will be limited. The more generic terms that could have retrieved additional relevant records (along with some non-relevant records) will not have been employed.

IV-C-v-4-b. Maximize Signal

The second case results from relaxed resource constraints. In this case, the objective is to maximize the number of records retrieved subject to a threshold signal-to-noise ratio. The general strategy for multiple iteration query development is to focus the initial iterations on expanding the number of relevant records retrieved, including the addition of non-relevant records, and then devote the last iteration mainly to eliminating the non-relevant records. A two iteration query development will be used to illuminate the concept.

The strategy for the first iteration of a two iteration signal maximization query is to generate a test query (initial guess), categorize the retrieved records into relevant and non-relevant bins, apply computational linguistics to each bin, and select only those phrases and phrase combinations that are moderately to strongly characteristic of the relevant bin for the modified query. The resulting records retrieved with this modified query will have a modest signal-to-noise ratio.

However, their coverage will be expanded relative to the previous (single iteration) case. The more generic terms that could retrieve additional relevant records (along with some non-relevant records) will have been employed.

The strategy for the second iteration of the two iteration signal maximization query is to use the modified query generated from the first iteration as a starting point, and categorize the retrieved records into relevant and non-relevant bins. Then, apply computational linguistics to each bin, and select mainly those phrases and phrase combinations that are strongly characteristic of the non-relevant bin for the modified query. Since new phrases will have resulted from the expanded relevant records retrieved by the modified first iteration query, some phrases and phrase combinations that are very strongly characteristic of the relevant bin can also be added. Again, the threshold ratio for phrase selection cutoff will be determined by the marginal utility of each phrase as a query term. Add these mainly negation phrases to the second iteration starting point query to produce the final modified query. The resulting records retrieved with this final modified query will have a very high signal-to-noise ratio, as confirmed by sampling a relatively few records retrieved by this final query, and their coverage will be expanded relative to the previous case.

In the truly resource unlimited case where the number of iterations could be relatively unbounded, the following approach would be taken. The number of relevant records after each iteration would be plotted as a function of iteration number, and the process would be terminated as the slope decreased to zero.

IV-D. Generate Test Query and Retrieve Records from Database; Select Sample

An initial guess of relevant query terms is made, and this test query is entered into the search engine. As in most iterative solutions, the iteration efficiency and speed will depend on the initial test query, although the final query structure will be moderately independent of the test query. As resources become more constrained, and the number of iterations is reduced, the final query quality becomes more dependent on initial test query quality.

IV-D-i. Criteria for Database Field Selection

As stated above, the query selection is also database, and database field, dependent. Since multiple databases with multiple fields tend to be used in many TDM studies, in theory a query would have to be tailored for each field in each database. In order to minimize multiple query development, the query

development philosophy with respect to databases and database fields has been the following. Select the database and field for query construction that will require development of the most complex query, and use this query, or segments thereof, to search all the fields in all the databases. This approach contains the inherent assumption that queries adequate for all the databases examined will be subsets of the most complex query developed.

For the semi-structured R&D databases that the author has used in developing the information retrieval process, experience has shown that the SCI tends to require the most complex query, since the language of research is quite diverse and filled with multiple meanings and interpretations. Within the SCI, the Abstract field contains the richest diversity and interpretation of language. Most of the queries used so far in the DT studies have been developed for the Abstract field of the SCI, and have been applied in whole or part to the other text fields in the SCI and the other semi-structured databases used. These SCI Abstract-based queries appear to have been adequate for high quality record retrieval from the other databases, for the topical areas examined so far.

IV-D-ii. Criteria for Sample Size Selection

After the test query, database, and database fields have been selected, the query is entered into the database search engine, and the resulting records are retrieved. Because of the potentially large volume of records that can be retrieved, the operational philosophy of the text mining studies using Simulated Nucleation has been to select a sample S0 of the records retrieved, and use this sample as the basis for analysis during each iterative step. The full complement of records retrieved is accessed when the final query has been entered into the search engine, and a small sample Sf has been examined to insure that the signal-to-noise ratio is at or greater than a desired threshold.

This sample size S0 depends on a number of factors, including which of the two Simulated Nucleation options is chosen. Since the reading of some or all of these records is a necessary component of the manually intensive Simulated Nucleation protocol, and since this reading is a time consuming process, the number of records selected for reading becomes a critical factor in the labor intensity of the process. For the semi-automated clustering option, minimizing the sample size is not as critical as in the manually intensive option.

From a statistical perspective, there are two main determinants of sample size S0. The weaker determinant is statistical adequacy for dividing the sample into two

categories: relevant and non-relevant. Relatively few records are required for making this black-white decision. The more stringent determinant is that the phrase and phrase combination frequencies resulting from the computational linguistics analysis of the sample are of sufficient magnitude that:

1) Important and unique phrases and phrase combinations can be separated from background phrases and phrase combinations within the same relevance/ non-relevance category, and
2) Important and unique phrases and phrase combinations in one relevance/ non-relevance category can be distinguished from phrases and phrase combinations in the other relevance/ non-relevance category to establish the dominant category for specific phrases and phrase combinations.

Because of language breadth and richness differences among myriad disciplines, fixed size samples can yield far different results in absolute values of phrase frequencies from the retrieved records. For example, Table A1 is a compendium of the statistics of high frequency technical content phrases from previous text mining studies using variants of Simulated Nucleation. The first column is the abbreviated title of the study. The second column is the number of records retrieved from the database (SCI) by the final query. The third column in the upper table is the frequency of occurrence of the highest technical content single word phrase, and the fourth and fifth columns are the frequencies for the commensurate double and triple word phrases. The third and succeeding columns in the lower table are the unitized version of the upper table; i.e., frequencies divided by number of records. The sixth column in the upper table is the frequency of occurrence of the second highest technical content single word phrase, and the seventh and eighth columns are the same for double and triple word phrases.

A scan of the unitized results shows that, for a specific phrase word length, the variation across different studies can be a factor of five or more. For a specific study, the single word phrases are about an order of magnitude higher frequency than the double word phrases, which are in turn typically factors of two to three larger than the triple word phrases. For some of the text mining studies in process, where the phrase combination frequencies are being tracked, the highest technical content phrase combinations have single word phrases for each member. The phrase combinations have frequencies typically between the frequencies of single and double word phrases not in combination.

Table A1 – Statistics of High Frequency Technical Content Phrases

STATISTICAL SAMPLING OF RETRIEVED RECORDS

TOPIC	#REC	TOP 1 WORD FREQ	TOP 2 WORD FREQ	TOP 3 WORD FREQ	SECOND 1 WORD FREQ	SECOND 2 WORD FREQ	SECOND 3 WORD FREQ
RIA	2300	1189	152	36	632	54	31
JACS	2150	1190	356	57	710	165	53
NES	5481	6517	579	227	6268	541	193
HYP	1284	3686	696	187	2061	517	69
AIR	4346	3959	329	176	3147	231	114
HYD	4608	5503	1839	393	3483	803	213
FUL	10515	12879	764	764	6791	479	191

NORMALIZED PHRASE FREQUENCIES PER RECORD RETRIEVED

TOPIC	TOP 1 WORD FREQ	TOP 2 WORD FREQ	TOP 3 WORD FREQ	SECOND 1 WORD FREQ	SECOND 2 WORD FREQ	SECOND 3 WORD FREQ
RIA	0.516957	0.066087	0.015652	0.274783	0.023478	0.013478
JACS	0.553488	0.165581	0.026512	0.330233	0.076744	0.024651
NES	1.189017	0.105638	0.041416	1.143587	0.098705	0.035213
HYP	2.870717	0.542056	0.145639	1.60514	0.402648	0.053738
AIR	0.910953	0.075702	0.040497	0.724114	0.053152	0.026231
HYD	1.194227	0.399089	0.085286	0.755859	0.174262	0.046224
FUL	1.224822	0.072658	0.072658	0.645839	0.045554	0.018165

Thus, the driving factor for statistical representation is the frequency of triple word phrases. Operationally, it is desirable if this frequency is about an order of magnitude higher than background. This is needed both for purposes of discrimination, and because frequencies of all types of phrases decline very rapidly from the maximum. Because of the wide range of frequencies possible, sample sizes in previous Simulated Nucleation studies have tended to be over-selected to insure statistically significant representation. In the future, as the algorithms become more efficient, dynamic sampling will be used. In the near future, frequencies will be examined for every sample group of 500, to ascertain if samples of 500, or 1000, will be adequate. In the far future, continuous sampling will be employed. Alternatively, if the clustering option produces category discrimination (relevant/ non-relevant) of equal fidelity to the manually intensive option, then the larger sample sizes can be easily obtained.

IV-E. Divide Sample Records into Relevant and Non-Relevant Categories

This is a crucial step in the process, since it represents the conversion of the study's scope from a theoretical abstract to an implementation requiring decisions

on each retrieved record. In the manually intensive option, the technical domain experts associated with the study read each of the sample records, and assign a level of relevance to each record. There are two main determinants to the quality of the relevance judgement: the quality of the records, and the expertise and background of the experts.

The quality of the records relative to the requirements for the relevance judgement task depends on the records' fields accessed and displayed for the relevance judgement, and the quality of the textual content contained within the fields accessed. There is a tradeoff of evaluation time vs. level of description for each field, or combination of fields, selected for the relevance judgement. If, for example, the only textual field supplied from each record for the relevance judgement were Keywords, the evaluation time per record will be short, but relatively little technical content and contextual information would be available to serve as a basis for making a credible relevance judgement. Addition of the Title field provides more context, with a negligible addition in evaluation time. Addition of the Abstract field provides substantially more technical content and context, along with a noticeable increase in evaluation time. Finally, addition of the full paper's text provides the most content and context, along with an unacceptably large increase in evaluation time. Most of the semi-structured databases supply the Abstract, as opposed to the full text of the paper, and a substantial additional amount of time would be necessary in the process of obtaining the full paper.

Thus far, the DT studies have used Keywords, Titles, and Abstracts as the text fields for making the relevance judgement. One operational problem experienced is that the quality of the Abstracts can vary substantially, and therefore can provide an uneven playing field for the relevance decision. While the medical literature appears to contain an informal protocol for the structure of its papers' Abstracts (Kostoff and Hartley, 2001k), and adherence to this protocol provides a threshold amount of useful information for relevance judgement purposes, the non-medical literature has no obvious structure or information threshold requirements. In some cases, the information in the Abstract is sufficiently vague that a credible relevance judgement cannot be made, and this degrades the quality of the final query. This Abstract quality problem results directly from the journal-database system being developer-driven rather than customer-driven, has been addressed elsewhere (Kostoff, 2003a, 2005i), and will not be discussed further.

Up to the present, the DT studies performed have used two relevance categories: relevant to the study's scope, and non-relevant to the scope. There is no reason

why degrees of relevance could not be used, and this approach may be implemented at some future time, but the two categories have proven adequate for past studies.

Experience has shown that the process of having to make a binary decision (on the relevance or non-relevance of a retrieved record) sharpens the focus of the study measurably. In addition, the process of reading a representative sample of Abstracts provides an excellent overview and substantial insights into the total discipline being examined. Therefore, the extra time spent by the experts on this step due to the over-sampling of the retrieved records is time well spent. For the TDM example, and many of the other studies as well, about 60 records per hour are processed when the full Abstracts are read, and perhaps 80-90 records per hour can be processed when the Abstracts are not read completely. Approximately 20-25 hours were required to read the records and place them into the appropriate category for the iterative step of the TDM study described here.

For many users, this will obviously be an overly time-consuming step. Present efforts are focused on reducing the number of iterations, and the number of sample records that have to be read for each iteration. In fact, use of the new algorithms for query term selection (to be described later) has approximately halved the number of iterations required, due to the heightened surgical precision of query term identification now possible with the aid of these algorithms.

The time consumption of this categorization step in the manually intensive option is the major driver for developing the clustering option. There will still be some sampling required for the clustering option, to assess the fidelity of the separation process for each database and thematic topic used. In addition, the clustering will not surmount the problems of vague Abstracts with minimal useful information, or Abstracts that border on the relevant/ non-relevant boundary. In general, the clustering process does separate the relevant from non-relevant records. The main reason for this separation is that the relevant records have multiple phrases/ words in common due to co-occurrence of similar phrases related to the main theme, while the non-relevant records tend to have only one multiple-meaning phrase/ word in common.

IV-F. Perform Computational Linguistics on Each Category

IV-F-i. Document Characteristics for Identifying Related Documents

Once the documents have been divided into relevant/ non-relevant categories (or any gradations of relevance that may be used in the future), then characteristics of records in each category can be obtained by a variety of computer-based techniques (bibliometrics, computational linguistics), and these characteristics can then be used to select other documents from the source database with similar characteristics. The underlying assumption is that records in the source database (e.g., SCI, EC) that have the same characteristics as the relevant records from the sample will also be relevant (or, more correctly, will have a high probability of being relevant), and records in the source database having the same characteristics as the non-relevant records from the sample will also be non-relevant. Under the broad umbrella of relevance, different degrees of relevance are of potential interest, depending on the overall study's objectives. Highly relevant, or similar, articles will provide comprehensive retrieval of papers in the specific target field of interest. Less similar articles, but still containing some similar characteristic features, will offer the opportunity for retrieval of papers in highly disparate, yet indirectly linked, disciplines. These types of papers offer the possibility of radical discovery and innovation from complementary literatures (Kostoff, 1999b, 2001d, 2001e, 2001i, 2002d, 2003b, 2003k, 2003q, 2004b, 2004r).

The myriad characteristics that can be used in the search for congruency depend on the breadth of features (fields) available in the source database search engine. In addition to the text fields in the semi-structured databases to which the author has applied computational linguistics for characteristic pattern matching, the author has used the following other fields in selected cases:

Authors; Journals; Institutions; Sponsors; and Citations.

Use of these fields to help identify relevant records, in addition to use of the text fields only, produces more relevant records than use of the linguistics patterns in the text fields alone. For TDM analyses whose objective is to provide an overview of a topical domain, and focus on trends and higher-order statistics, the computational linguistics will result in more than adequate statistically representative samples of retrieved records. For TDM analyses whose objective is to impact organizational operations and specific funding decisions, as many of the above fields as is practical should be used to identify as many relevant records as possible.

The specific rationale for using some of these other fields is described briefly.

IV-F-i-a. Author Field

An author of a few relevant documents will tend to work in technical areas similar to those characteristic of the relevant documents. Therefore, a search for other publications by the same author will have good probability of retrieving similar relevant documents. One problem with using the Author field is that present-day semi-structured databases don't assign unique names or numbers to each author. Searching for publications from an author with a common name could result in retrieval of many extraneous records. If manual filtering is performed, it would require a time-intensive filtering process. Cluster filtering, if it proves feasible for high fidelity results, would reduce the time intensity of the separation process. Again, this problem results from the technical community depending on developer-driven databases rather than customer-driven databases, has been addressed elsewhere (Kostoff, 2003a, 2005i), and will not be discussed further.

IV-F-i-b. Journal Field

A journal found to contain a few relevant documents will probably contain many more, given the specialized nature of most journals.

IV-F-i-c. Institution Field

An institution that produces a few relevant documents could be expected to produce many others as well. Institutions tend to concentrate their efforts in core competency areas. Accession to these institutions' program outputs could result in uncovering related documents. Unfortunately, institution organizational unit levels specified by the author, and institutional abbreviations are not standardized. As in the author field case, either substantial manually intensive filtering is required, or the problem may be alleviated if cluster filtering proves successful.

IV-F-i-d. Sponsor Field

An S&T sponsor whose output includes a few relevant papers, or more specifically a program or project from such a sponsor, could be expected to produce other relevant documents. Sponsors, like institutions, tend to concentrate their funds in core competency areas. Accession to these sponsors' program outputs could result in uncovering related documents. One problem is that this sponsor database field appears only sporadically in semi-structured R&D databases. Again, this problem results from the technical community depending on developer-driven databases rather than customer-driven databases, has been addressed elsewhere for the case of the SCI (Kostoff, 2003a, 2005i), and will not be discussed further.

IV-F-i-e. Citation Field

There are at least three ways in which the citation field can be used to help identify additional relevant papers.

IV-F-i-e-1. Papers that Cite Relevant Documents

Papers that cite relevant documents tend to have thematic similarity to the relevant document. The more relevant documents cited by a given paper, the higher is the probability that the citing document will be relevant. One of the problems here is that cross-linked citations are not available in many semi-structured R&D databases. The SCI is the only database the author has used that contains this capability.

IV-F-i-e-2. Papers Referenced in Relevant Documents

In parallel with the previous sub-section, papers that are cited by relevant documents tend to be relevant. The more times a paper is cited by different relevant documents, the higher is the probability that the cited paper will be relevant.

IV-F-i-e-3. Other Papers Cited by Paper that Cites Relevant Documents

The first two examples dealt with relevance resulting from direct citations, with the probability of relevance increasing as the numbers of citations increased. This third example is one step removed from a direct relationship. A paper has increased chances of being relevant if it is cited by a paper that also cites relevant documents. The larger the number of relevant documents that the citing paper references, and the larger the number of citing papers that reference the paper of interest and also cite other relevant papers, the higher is the probability that the target paper will itself be relevant.

In addition, other papers by authors/ journals/ organizations that cite relevant papers have increased probability of being relevant, as well as other papers/ journals/ organizations that are cited by relevant papers. The reasons parallel those given above for authors, journals, and organizations.

The remainder of this section will focus on use of the text fields as a source of linguistic patterns for identifying related documents.

IV-F-ii. Linguistics Patterns for Identifying Related Documents

The purpose of this step is to identify linguistic patterns uniquely characteristic of each category (relevant and non-relevant records), and use this information to modify the query. The underlying assumption is that records in the source database (e.g., SCI, EC) that have the same linguistic patterns as the relevant records from the sample will also be relevant (or, more correctly, will have a high probability of being relevant), and records in the source database having the same linguistic patterns as the non-relevant records from the sample will also be non-relevant. Linguistic patterns characteristic of the relevant records would be used to modify the query such that additional relevant records would be retrieved from the source database. Linguistic patterns characteristic of the non-relevant records would be used to modify the query such that existing and additional non-relevant records would not be retrieved.

Semantics and syntax of text are extremely complicated, and many types of linguistic patterns can be selected for characterizing records. DT has focused on two types of congruency metric patterns for identifying candidate query modification terms: phrase frequencies and phrase proximity statistics. In entropic terms, these pattern metrics are macro-state descriptors. There are many linguistic micro-states (where the ordering of the phrases is included in the descriptors) that correspond to any one macro-state. These two high entropy macro characteristics have proven to be adequate for identifying the full complement of relevant records, although conceivably in the future more micro-state oriented metrics may be added as well to provide lower entropy measures of congruency.

For the TDM study, the frequencies of all single, adjacent double, and adjacent triple word phrases in the Abstract in each category were obtained with the DT algorithms, and then the phrases in close proximity to selected theme phrases in the Abstracts were also obtained with the DT algorithms. The next few paragraphs summarize how the phrases and phrase combinations actually used for query modification were obtained from the raw computer output of this step using the recently developed selection-support algorithms.

In the following paragraphs, construction of a query with the aid of the new algorithms is presented as a two-step process. At the beginning of the first step (described in detail in the next section IV-G), all high technical content phrases occurring in the total sample are listed in descending numerical order based on

their frequency of occurrence in each category. Then, four types of generic judgements are made on each phrase.

1) It belongs in the modified query as a stand-alone phrase (e.g., the phrase 'TEXT DATA MINING' belongs in the modified query as is)
2) Its components may belong in the modified query in some combination (e.g., the combination 'TEXT' and 'DATA MINING' should be added to the query)
3) The phrase and any component permutations do not belong in the modified query (e.g., 'TEXT DATA MINING' or any combination of component phrases does not belong in the modified query)
4) It may be a candidate for the modified query in combination with some other phrase(s) (e.g., 'TEXT DATA MINING' could be too generic to be added to the query as a stand-alone phrase, but could gain enough specificity if added to the query in combination with another phrase).

As stated before, if a large number of iterations are used to construct the query, the terms added to the query in the early iterations will be those characteristic of the relevant category (for expanding records retrieved). The (negation) terms added to the query in the last iteration will be those characteristic of the non-relevant category (for contracting records retrieved). The component of query construction resulting from the first step is the group of phrases positively identified from the first type of judgement.

In the second step (described in section IV-H), proximity runs are made using the candidate phrases from the fourth type of judgement (above) as themes. The resulting phrase combinations are listed in descending numerical order based on their frequency of occurrence in each category of the total sample. Then, two types of judgements are made on each phrase combination.

1) It belongs in the modified query as a phrase combination
2) It does not belong in the modified query

While the fourth type of judgement (above) is not presently made for these binary phrase combinations in the second step, it could be used in the future if it is decided to incorporate higher than binary combinations in the query.

The judgements in each step are made within the context of the larger query development objectives, detailed in section IV-A, and summarized as follows: The over-riding objective of query construction is to 1) select the minimum number of phrases that will 2) retrieve the maximum number of relevant records with 3) the

requisite threshold signal-to-noise ratio. Specifically, when dealing with a sample of records, the objective is to select the minimum number of phrases that would retrieve the maximum number of sample records in each of the two categories. The assumption is that this retrieval efficiency for the sample would extrapolate to the total records retrieved.

IV-G. Use New Algorithms to Identify Phrases Unique to Each Category

IV-G-i. Structure of Phrase Selection Algorithm

Phrases and their associated frequencies are extracted from the text of the records in the relevant and non-relevant categories. These phrases and frequencies are then imported into the ACCESS-based database, hereafter referred to as a template. Then, a normalization procedure is performed on the frequencies such that they would represent the situation where the numbers of records in each category are equated.

The template has two major components, phrase frequency and phrase proximity (phrase combination frequency). Each has a separate display window, but both components are linked algorithmically. The phrase frequency component aids in the selection of stand-alone phrases for the modified query and potential anchor phrases for phrase combinations. The phrase proximity component aids in selection of phrase combinations for the modified query. The present section describes the phrase frequency component.

Figure A1 shows the display window from the template's phrase frequency component for the TDM example. Each template row contains a phrase extracted from the Abstracts' text, and nine associated numerical and other indicators for that phrase. The analyst sees the phrase and its indicators in ten fields/ columns on the computer screen. Six of the fields/ columns are shown in Figure A1. Proceeding from the leftmost column, the columns/ fields are defined as:

FIGURE A1

PHRASE	NORM REL FREQ	NORM NON-REL FREQ	RATIO	DOM FREQ	DOM CATEG
IR	207	3	69	207	Relevant
TEXTUAL	68	3	23	68	Relevant
SEARCH ENGINES	243	11	22	243	Relevant
DOCUMENT	331	16	21	331	Relevant
ENGINES	279	16	17	279	Relevant

RELEVANCE	286	19	15	286	Relevant
SPATIAL	7	107	15	107	Non-Relevant
DOCUMENTS	427	32	13	427	Relevant
LEXICAL	65	5	13	65	Relevant
RELEVANCE FEEDBACK	47	4	12	47	Relevant

1) Shown in Figure A1

*PHRASE – Presently, this entry is a single, adjacent double, or adjacent triple word phrase that was extracted from one or both of the relevant/ non-relevant categories. It survived a filtering by a trivial phrase algorithm, and its frequency of occurrence in either the relevant or non-relevant category was above some pre-defined threshold.

There are intrinsically two types of phrases: those phrases included in the initial iteration test query, and those phrases not included in the test query, but extracted from the records retrieved with use of the test query. The difference between these two types of phrases is significant with respect to the interpretation and utilization of the associated numerical indices shown, as discussed in the definitions of the next three fields.

*NORMALIZED RELEVANT FREQUENCY – The occurrence frequency of the phrase in the relevant category, after the normalization has been done on the category. If the phrase was included in the test query, the frequency represents not only the relative sample occurrence, but also the expected relative occurrence in the total source database. If the phrase was not included in the test query, the frequency still represents the relative sample occurrence, but may not be a good indicator of the expected relative occurrence in the total source database.

*NORMALIZED NON-RELEVANT FREQUENCY – The occurrence frequency of the phrase in the non-relevant category, after the normalization has been done on the category. The same argument about the significance of the phrase's appearance in the test query used in the previous paragraph holds here as well.

*RATIO – The ratio of the above two normalized frequencies, with the dominant frequency selected for the numerator. The same argument about the significance of the phrase's appearance in the test query used in the previous two paragraphs holds here as well. This metric is used as the starting point for selecting candidate query terms, with the single iteration query limited to the higher metric values, and the maximal coverage multi-iteration query incorporating somewhat lower metric

values as well. The ability to compute this metric is one of the upgrades to Simulated Nucleation.

*DOMINANT FREQUENCY – The larger of the normalized frequencies of the phrase in the relevant or non-relevant categories.

*DOMINANT CATEGORY – The category in which the phrase has the larger normalized frequency.

2) Not shown in Figure A1

*INCLUDE – This field is a block that the analyst checks if he/ she decides the phrase (e.g., SEARCH ENGINES) is a stand-alone candidate for the modified query.

*INCLUDED PHRASE – This field is a block that is automatically checked if the phrase in the first column (e.g., WEB SEARCH ENGINES) includes a more generic phrase (e.g., SEARCH ENGINES) that received a check in the INCLUDE field. The purpose of this field is to eliminate duplications, in order to satisfy the query development criterion of minimal number of query terms.

In the preceding example, use of the phrase SEARCH ENGINES in the query will automatically retrieve all of the records that contain the more specific phrase WEB SEARCH ENGINES. Therefore, there is no need for the analyst to consider WEB SEARCH ENGINES once the phrase SEARCH ENGINES has been selected. As will be shown later, in the section on phrase combinations, the INCLUDED PHRASE field in the phrase combination template is automatically checked if the phrase combination on a template row (e.g., DATA and SEARCH ENGINES) includes a phrase (e.g., SEARCH ENGINES) that received a check in the INCLUDE field.

This phrase tracking capability is one of the upgrades to Simulated Nucleation. It allows the analyst to eliminate duplicative phrases and phrase combinations without having to remember which parent phrases or phrase combinations had been selected previously. Without this capability, examination of the many thousands of candidate phrases and phrase combinations that occur with use of Simulated Nucleation, and identification of those that are not duplicative, are not operationally feasible.

*THEME CANDIDATE– This field is a block that the analyst checks if he/ she decides the phrase (e.g., DATA) could be a candidate for a phrase combination (e.g., DATA and MINING) in the modified query.

*ADDITIONAL NEW RECORDS – This field informs the analyst of the number of additional sample records that the phrase in the first column would retrieve. The purpose of this field is to eliminate effective duplications resulting from co-occurrence, in order to satisfy the query development criterion of selecting the minimum number of phrases that would retrieve the maximum number of sample records in each of the two categories. There are seven other fields used (not reported here) that provide a full accounting of how the candidate phrase is distributed within the sample records (distribution in relevant and non-relevant sample records, cumulative and marginal distributions, etc). This marginal utility capability is one of the upgrades to Simulated Nucleation, and is described in more detail in section IV-I.

IV-G-ii. Use of Phrase Selection Algorithm

To facilitate the initial phrase selection judgements, the phrases are sorted by the ratio of frequencies (column 4), in decreasing order. The higher ratio phrases are more uniquely characteristic of a specific category. From Figure A1, TEXTUAL is more uniquely characteristic of relevant records in the sample, whereas SPATIAL is more uniquely characteristic of non-relevant records in the sample. For sample categories of equal numbers of records, TEXTUAL appears 23 times more frequently in relevant records than non-relevant records, whereas SPATIAL appears 15 times more frequently in non-relevant records than relevant records.

IV-G-ii-1. Minimize Iterations

In selecting candidate phrases for the minimal iteration query, the technical expert(s) employs the following protocol.

Start at the top of the list (highest ratio). If the field INCLUDED PHRASE has a check, go to the next phrase. If the field INCLUDED PHRASE does not have a check, examine the dominant category.

If the phrase:

1) is dominant non-relevant;
2) has a high marginal utility based on the sample;

3) has reasons for its appearance in the non-relevant records that are well understood; and
4) IS PROJECTED TO ELIMINATE RECORDS FROM THE SOURCE DATABASE (E.G., SCI, EC) MAINLY NON-RELEVANT TO THE SCOPE OF THE STUDY (especially important in the later iteration steps),

then select it as a candidate stand-alone query modification phrase (i.e., enter a check in the INCLUDE block).

If the phrase:

1) is dominant relevant;
2) has a high marginal utility based on the sample
3) has reasons for its appearance in the relevant records that are well understood; and
4) AND IS PROJECTED TO RETRIEVE ADDITIONAL RECORDS FROM THE SOURCE DATABASE (E.G., SCI) MAINLY RELEVANT TO THE SCOPE OF THE STUDY (especially important in the earlier iteration steps),

then select it as a candidate stand-alone query modification phrase (i.e., enter a check in the INCLUDE block). If these four criteria are not met, do not select the phrase as a stand-alone query modification candidate. If these four criteria are met, and the phrase contains multiple words, view the phrase selection as tentative. As the next section on phrase combinations shows, there may be combinations of the phrase's component words that are uniquely characteristic of one of the categories. Use of the phrase's component words (e.g., SEARCH and ENGINES) instead of the actual phrase (e.g., SEARCH ENGINES) would retrieve more desired records, and therefore the combination of the phrase's component words would be used instead of the actual phrase.

The first two of the criteria above (dominant category, high marginal utility) are numerically based and straight-forward. The third criterion (understand appearance in dominant category) is essentially a requirement for, and supportive of, the fourth criterion (project dominant category occurrence in total source database). For the first type of phrase discussed previously (included in test query), the source database projection is straight-forward, and is reflected by the ratio metric. For the second type of phrase discussed previously (not included in test query), the actual source database occurrence ratio may be far different from the projection based on the ratio metric. The IR example discussed after the next paragraph is an excellent demonstration of the mis-estimate of total source

database occurrence possible with the second type of phrase. This estimation error for the second type of phrase is reduced as the third criterion is met more stringently.

The few text mining studies that have been done with these latest algorithmic capabilities for the minimal (one) iteration case show that if high ratio dominant relevant terms are selected (with care), essentially all the retrieved records are relevant, and dominant non-relevant terms are not required. Examples from the general minimal iteration case of the TDM study will now be demonstrated.

As an example from Figure A1, the phrase IR (an abbreviation used in many of the TDM study Abstracts for information retrieval) is dominant relevant (ratio of 69) based on the sample, and has a high marginal utility based on the sample. However, it is not 'projected to retrieve additional records from the source database mainly relevant to the scope of the study'. A test query of IR in the source SCI database showed that IR occurred in 65740 records dating back to 1973. Examination of only the first thirty of these records showed that IR is used in science and technology as an abbreviation for InfraRed (physics), Immuno-Reactivity (biology), Ischemia-Reperfusion (medicine), current(I) x resistance(R) (electronics), and Isovolume Relaxation (medical imaging). The number of records in this database in which IR occurs as an abbreviation for information retrieval is probably one percent of the total records retrieved containing IR, or less. Therefore, the phrase IR is not selected as a stand-alone query modification candidate.

Continuing on Figure A1, the phrase SEARCH ENGINES is dominant relevant based on the sample, has a high marginal utility based on the sample, tends to occur in Abstracts focused on the information retrieval component of textual data mining, and is 'projected to retrieve additional records from the source database mainly relevant to the scope of the study'. Therefore, the phrase SEARCH ENGINES is selected as a stand-alone query modification candidate, and a check is entered in its INCLUDE block.

Continuing further on Figure A1, the phrase SPATIAL is dominant non-relevant, has a high marginal utility based on the sample, tends to occur in Abstracts focused on numerical data mining, and is 'projected to eliminate records from the source database mainly non-relevant to the scope of the study'. Whether SPATIAL is selected as a candidate stand-alone query term depends on the strategy for including or excluding terms from the original test query.

If the terms from the test query are retained for the modified query, and terms identified from the computational linguistics results added to this query, then SPATIAL is selected as a candidate stand-alone query modification phrase, and a check is entered in its INCLUDE block. If the terms from the test query are in general not retained as a starting point for the modified query, and only terms identified from the computational linguistics results are used to construct the modified query (some of these terms could also have been in the test query), then SPATIAL is not selected as a candidate stand-alone query modification phrase. The reasoning is straight-forward: SPATIAL was selected because it occurred mainly in non-relevant records that resulted from the inclusion of some of the terms in the test query. If these non-relevant record generating terms are not required for the modified query, than there is no reason to use terms that would negate these (non-existing) non-relevant records.

As these examples from Figure A1 show, substantial judgement must be exercised when selecting candidate phrases, even when using this new phrase selection-support algorithm. When potentially dominant relevant query modification terms are being evaluated, one has to consider whether substantial amounts of non-relevant records will also be retrieved, and when potentially dominant non-relevant query modification terms are being evaluated, one has to consider whether substantial amounts of relevant records will not be retrieved. For a fixed number of query modification iterations, excess 'noise' records retrieved by broad query terms with multiple meanings will degrade the overall quality of the retrieved record database. Conversely, if the constraint is a fixed signal-to-noise ratio for the retrieved records database, then additional iterations will be required to remove the 'noise' records resulting from the overly broad and multiple-meaning terms. This translates into additional time and other resources.

Thus, the relation of the candidate query term to the objectives of the study, and to the contents and scope of the total records in the full source database (i.e., all the records in the SCI, not just those retrieved by the test query), must be considered in query term selection. The quality of this selection procedure will depend upon the expert(s)' understanding of the scope of the study, and the expert(s)'understanding of the different possible meanings of the term across many different areas of R&D. *This strong dependence of the query term selection process on the overall study context and scope makes the 'automatic' query term selection processes reported in the published literature very suspect.*

Continuing the selection protocol, proceed down the list identifying candidate query terms until one of two conditions is reached. Either the number of terms

sums to some pre-determined maximum (e.g., a given search engine has a ceiling of fifty query terms), or the ratio of frequencies reaches a threshold. For the TDM study, and other very recent studies in which this algorithm was used, a phrase frequency ratio threshold of eight was used for the minimal iteration objective. In addition, for the TDM study, a condition of marginal utility (new 'signal' records retrieved divided by new 'noise' records retrieved) for each term was used. This condition was determined by the latest selection-support algorithm, and will be described in section IV-I.

IV-G-ii-2. Maximize Coverage

Only the differences in procedures between this case and the previous minimal (one) iteration case will be discussed. Assume the baseline minimal iteration case is where all the modified query terms result from the computational linguistics results. For this single iteration case, only the high ratio terms characteristic of the relevant records are selected as candidate stand-alone query modification terms.

For the maximal coverage case, the initial iterations are not restricted to selecting high ratio terms characteristic of the relevant records as stand-alone query modification terms. Lower ratio terms characteristic of the relevant records can be selected as well for the initial iterations. This strategy will result in the retrieval of more relevant records due to the use of the (typically) broader terms characteristic of the lower ratios, as well as the retrieval of some non-relevant records during the early iterations.

To eliminate the non-relevant records, there are two major options. In the more conservative option, the final iteration consists of identifying only the high ratio phrases characteristic of the non-relevant records, and adding them to the test query for the final iteration. Since all new records were added as a result of the query from the previous iteration, and no new records will be added as a result of adding negation terms to this query, there is no chance that new 'noise' records will be added as a result of the final query.

In the more risky option, the final iteration consists of identifying the high ratio phrases characteristic of the relevant and non-relevant records, and adding them to the test query. This would have two potentially negative effects. First, the phrases characteristic of the relevant records would be the more restricted high ratio phrases rather than the more inclusive moderate ratio phrases. This is necessary to increase the probability that new 'noise' records will be minimized. Second, the negation terms would strictly address the records retrieved by the query resulting

from the previous iteration, and one could not be completely sure that addition of high ratio phrases characteristic of the relevant records did not retrieve an anomolously large number of 'noise' records. On the positive side, one iteration has been eliminated by this option.

IV-H. Use New Algorithms to Identify Phrase Combinations Unique to Each Category

IV-H-i. Selection of Candidate Phrases for Phrase Combinations

Until this point in the query term selection protocol, only the high frequency ratio relatively specific phrases have been considered for the minimal iteration case, and moderate to high ratio relatively specific phrases for the maximal coverage case. Now, the addition of some lower frequency ratio more generic phrases to the process will be discussed. Phrases that have a high absolute frequency value in the dominant category, but a relatively low frequency ratio, could have the potential to be used in combination with other phrases to still retrieve (or eliminate) a significant number of records in the desired category. One objective of the following step is to identify those high frequency low ratio phrases that have the potential for such beneficial combinations. In addition, phrases that have a high absolute frequency value in the dominant category and a high frequency ratio, but are too generic to be used in a stand-alone mode, could have the potential to be used in combination with other phrases to still retrieve (or eliminate) a significant number of records in the desired category.

Continuing the protocol, the expert(s) re-sorts the rows in the ACCESS template's phrase frequency window by absolute frequency, first by dominant relevant category, then by dominant non-relevant category. Then, the analyst identifies perhaps a dozen of the highest frequency high and low ratio promising phrases, and enters a check into the THEME CANDIDATE block for each of these phrases. Again, judgement plays a very key role in this step, since the promising phrases should have high potential to anchor combinations that would be highly relevant to the study's scope (or highly non-relevant). Typically, the higher the specificity of a phrase, the higher will be its frequency ratio, and the more likely it will result in combinations that are uniquely characteristic to the appropriate category.

FIGURE A2

PHRASE	NORM REL FREQ	NORM NON-REL FREQ	RATIO	DOM FREQ	DOM CATEG
INFORMATION	1796	637	3	1796	Relevant
RETRIEVAL	895	245	4	895	Relevant
SEARCH	736	194	4	736	Relevant
SYSTEM	596	740	1	740	Non-Relevant
WEB	527	69	8	527	Relevant
KNOWLEDGE	457	1264	3	1264	Non-Relevant
INFORMATION RETRIEVAL	448	85	5	448	Relevant
DATA	436	1151	3	1151	Non-Relevant
SYSTEMS	436	486	1	486	Non-Relevant
DOCUMENTS	427	32	13	427	Relevant
PAPER	410	595	1	595	Non-Relevant

Figure A2 shows a template phrase frequency window in which the rows are sorted by absolute (normalized) frequency of relevant category-dominant phrases, in descending order. The phrase INFORMATION has high absolute frequency of occurrence in both categories, and a modest focus. Its frequency ratio (~3) offers promise that probably many phrases could be located that would form a combination with INFORMATION strongly characteristic of relevant records. INFORMATION is therefore perceived to have high potential to anchor combinations that would appear in substantial numbers of records strongly relevant to the study's scope. Because of the large value of absolute frequency in both categories, INFORMATION may also have potential to anchor combinations that would appear in reasonable numbers of records strongly non-relevant to the study's scope as well. The probability of anchoring combinations characteristic of relevant records would be greater than the corresponding probability of anchoring combinations characteristic of non-relevant records. Therefore, INFORMATION is selected as a candidate query phrase combination anchor, and a check is entered in its CANDIDATE THEME block.

Continuing on Figure A2, the phrase SYSTEM also has high absolute frequency of occurrence in both categories. However, its focus is weak. Its frequency ratio (~1) is sufficiently low that the relative probability is low that many phrases could be located that would form a combination with SYSTEM strongly characteristic of either relevant or non-relevant records. Therefore, SYSTEM is not selected as a candidate query phrase combination anchor.

FIGURE A3

PHRASE	FREQ NORM REL	REL FREQ NORM NON-	RATIO	FREQ DOM	CATEG DOM
KNOWLEDGE	457	1264	3	1264	Non-Relevant
DATA	436	1151	3	1151	Non-Relevant
SYSTEM	596	740	1	740	Non-Relevant
INFORMATION	1796	637	3	1796	Relevant
PAPER	410	595	1	595	Non-Relevant
SYSTEMS	436	486	1	486	Non-Relevant
MODEL	263	441	2	441	Non-Relevant

Figure A3 shows a template phrase frequency window in which the rows are sorted by absolute frequency of non-relevant category-dominant phrases, in descending order. The phrase KNOWLEDGE has high absolute frequency of occurrence in both categories, and a modest focus. Its frequency ratio (~3) offers promise that probably many phrases could be located that would form a combination with KNOWLEDGE strongly characteristic of non-relevant records. KNOWLEDGE is therefore perceived to have high potential to anchor combinations that would appear in substantial numbers of records strongly non-relevant to the study's scope. Because of the large value of absolute frequency in both categories, KNOWLEDGE may also have potential to anchor combinations that would appear in reasonable numbers of records strongly relevant to the study's scope as well. The probability of anchoring combinations characteristic of non-relevant records would be greater than the corresponding probability of anchoring combinations characteristic of relevant records. Therefore, KNOWLEDGE is selected as a candidate query phrase combination anchor, and a check is entered in its CANDIDATE THEME block.

Continuing on Figure A3, the phrase PAPER also has high absolute frequency of occurrence in both categories. However, its focus is weak. Its frequency ratio (~1) is sufficiently low that the relative probability is low that many phrases could be located that would form a combination with PAPER strongly characteristic of either relevant or non-relevant records. Therefore, PAPER is not selected as a candidate query phrase combination anchor.

Each of these selected high frequency high and low ratio phrases is entered into the DT phrase proximity algorithm. Phrases in the sample's aggregated records (Abstracts, in recent studies) that are located in close proximity to each high frequency low ratio phrase (essentially located in the same Abstract), in both the relevant and non-relevant categories, are identified.

This section ends with a caveat. To avoid division by zero and subsequent ratio of infinity, a phrase that has a finite frequency in one category (relevant or non-relevant) and does not appear in the other category is given a default frequency of one in the non-appearing category. Thus, a low ratio phrase could contain a substantial number of noise records if it also has high absolute occurrence frequency, but could have no noise records if its absolute occurrence frequency is very low. The absolute occurrence frequency of a low ratio phrase should be considered when deciding how the phrase should be used in the analysis.

IV-H-ii. Structure of Phrase Combination Selection Algorithm

These phrase combinations and their associated frequencies are extracted from the text of the records in the relevant and non-relevant categories. These phrase combinations and frequencies are then imported into the ACCESS-based template. Then, a normalization procedure is performed on the phrase combination frequencies, similar to that performed on the phrase frequencies.

Figure A4 shows the display window from the template's phrase combination frequency component for the TDM example. Each template row contains a phrase combination extracted from the Abstracts' text, and nine associated numerical and other indicators for that phrase. The analyst sees the phrase combination and its indicators in ten fields/ columns on the computer screen, and seven of the fields/ columns are shown in Figure A4.

FIGURE A4

THEME	PHRASE	NORM REL FREQ	NORM NON REL FREQ	RATIO	DOM FREQ	DOM CATEG
INFORMATION	DOCUMENT	224	1	224	224	Relevant
INFORMATION	IR	198	1	198	198	Relevant
RETRIEVAL	IR	184	1	184	184	Relevant
QUERY	SEARCH	182	1	182	182	Relevant
INFORMATION RETRIEVAL	IR	178	1	178	178	Relevant
SEARCH	SEARCH ENGINES	178	1	178	178	Relevant
INFORMATION RETRIEVAL	SEARCH	168	1	168	168	Relevant
KNOWLEDGE	USING	1	157	157	157	Non_Relevant
SEARCH	SYSTEMS	140	1	140	140	Relevant
RETRIEVAL	METHOD	134	1	134	134	Relevant

KNOWLEDGE	FUZZY	1	132	132	132	Non_Relevant
RETRIEVAL	KNOWLEDGE	126	1	126		Relevant

Proceeding from the leftmost column, the columns/ fields are defined as:

*THEME – This entry is a single, adjacent double, or adjacent triple word phrase that was identified as a promising phrase combination anchor from the stand-alone phrase selection process.

*PHRASE – This entry is a single, adjacent double, or adjacent triple word phrase that was physically located within a specified number of words from the theme phrase in one or both of the relevant/ non-relevant categories. The capability also exists to specify co-occurrence within the same Abstract, paragraph, or sentence. The phrase survived a filtering by a trivial phrase algorithm, and the frequency of its occurrence in combination with the theme phrase in either the relevant or non-relevant category in the aggregate sample was above some pre-defined threshold.

The remaining fields displayed have the same headings and definitions as those on the template's phrase frequency window, and will not be repeated. The remaining fields not displayed have the same headings and definitions as those on the template's phrase frequency window, with the exception that the THEME CANDIDATE field has been eliminated.

IV-H-iii. Use of Phrase Combination Selection Algorithm

The selection procedure for phrase combinations now proceeds the same as for stand-alone phrases shown previously, and the same type of logic and reasoning is used. Because of space limitations, no examples will be provided.

This phrase combination selection procedure tends to:

1) involve many more database entries to examine than the phrase only procedure;
2) have many more high ratio entries due to the increased specificity of the more detailed entries; and
3) have somewhat lower absolute frequency values due to the fact that higher specificity terms have reduced occurrence frequencies.

Obviously, this combination procedure could be continued for three phrases (a AND b AND c), four phrases, and so on. So far, the surgical precision provided by the two-phrase combination has been adequate for study purposes.

IV-I. Use New Algorithms to Identify Marginal Value of Adding Each Phrase and Phrase Combination to Query

IV-I-i. Manual Query Selection Procedure

Marginal utility, in the present context, is a measure of the ratio of 1) additional desirable records (signal) retrieved by the addition of a query term to 2) additional non-desirable records (noise) retrieved by the addition of this term. It is also used as an efficiency measure for eliminating undesirable records. It becomes an important consideration when query size reduction is required.

The frequency ratio metric would approximately reflect marginal utility only when phrases do not co-occur in the same Abstract. For the first few phrases selected, there is probably a relatively modest level of co-occurrence, because of the low-density factor. As more and more phrases are selected for query modification candidates, the number of un-retrieved sample records in which the next candidate phrase would appear decreases substantially. Thus, some method of taking co-occurrence into account is necessary for achieving the initial query development objective of selecting the minimal term query for maximum record retrieval.

The upgraded query term selection-support algorithm shows the aggregate level of co-occurrence, and allows the marginal utility of each additional query term to be estimated. All the sample records from each category (relevant and non-relevant) are entered in the ACCESS database. The records in which each query term appears are tracked continuously. When a candidate query modification term is selected, the number of new sample

records in the desired category (signal) in which the term appears is identified, as well as the number of new sample records in the un-desired category (noise). The aggregate, as well as the marginal, number of sample records in each category is tracked, allowing estimates of the marginal benefit of each term to the query. For consistency, the normalization used to balance relevant and non-relevant record categories is also employed to track marginal utility.

Figure A5 shows how the marginal utility would operate conceptually.

FIGURE A5 – MARGINAL UTILITY ESTIMATION

RECORD→	RELEVANT			NON-RELEVANT			RELEV	DELTA RELEV	DELTA NON RELEV	CUMUL RELEV	CUMUL NON
PHRASE/	R1	R2	R3	R4	R5	R6	R7				
P1	X	X						2	0	2	0
P2		X	X			X		1	1	3	1
P3			X	X		X	X	1	1	4	2
P4					X		X	1	0	5	2

In the first column, the Pi represent different candidate query phrases. The Rj column headings represent different records. An X in element ij means that phrase Pi is present in record Rj. Records R1-R5 have been judged to be relevant, and records R6-R7 are non-relevant. Thus, phrase P1 is present in the relevant records R1 and R2, and phrase P4 is present in the relevant record R5, and the non-relevant record R7.

The four columns on the right contain summary statistics for the marginal utility computation. The column headed DELTA RELEV contains the additional number of relevant records identified by the candidate query phrase. Thus, phrase P1 appears in the two previously unmarked relevant records R1 and R2, and the number 2 is entered into the DELTA RELEV column. Phrase P2 appears in relevant records R2 and R3, but since R2 contains the previously entered phrase P1, only one additional relevant

record (R3) has been identified by phrase P2. Therefore, the number 1 is entered into the column DELTA RELEV.

The column headed DELTA NON-RELEV contains the additional number of non-relevant records identified by the candidate query phrase. Thus, phrase P2 appears in the previously unmarked non-relevant record R6, and the number 1 is entered into the DELTA NON-RELEV column. Phrase P3 appears in non-relevant records R6 and R7, but since R6 contains the previously entered phrase P2, only one additional non-relevant record (R7) has been identified by phrase P3. Therefore, the number 1 is entered into the column DELTA NON-RELEV. Phrase P4 appears in the non-relevant record R7, but since R7 contains the previously entered phrase P3, no additional non-relevant records are identified by P4.

The columns headed CUMUL RELEV and CUMUL NON-RELEV contain running sums of the columns headed DELTA RELEV and DELTA NON-RELEV, respectively. A recently published study provides a detailed example of marginal utility application to query development (Kostoff et al, 2004a). For the specific problem studied, use of the first 100 terms of a (greater than) 150 term query resulted in no loss of fidelity of retrieved records.

IV-I-ii. Automated Query Selection Procedure

Presently, this procedure is being automated. Once a pool of candidate query modification terms has been selected, and the maximum number of query terms has been specified, the automation algorithm will examine each term in the pool, and the terms that provide the greatest marginal benefit at each step will be added to the query. Two approaches to computing 'greatest marginal benefit' will be described. The first has the objective function of maximizing signal-to-noise ratio of the records retrieved, and the second has the objective function of maximizing total relevant records retrieved subject to a signal-to-noise ratio floor.

IV-I-ii-a. Maximum Signal-to-Noise Ratio

The objective of this option is to maximize the signal-to-noise ratio (relevant/ non-relevant records) of the retrieved relevant records, subject to the constraint of an upper bound on the number of query terms allowed by the search engine. As an example of this case, suppose that 200 candidate

query terms have been identified by the expert(s). Suppose further that a query limit of fifty terms has been specified. Then, this semi-automated optimization protocol would proceed as follows.

1) The highest ratio candidate phrase is selected to initialize the system. Call this phrase Term 1. Its approximate marginal utility would be the ratio of frequencies, since there is no co-occurrence of selected phrases at this point.
2) Then, each of the remaining 199 terms is examined. The term with the highest marginal utility (Term 2) is identified. Term 2 and Term 1 are the selected query modification terms at this point.
3) Then, each of the remaining 198 terms is examined. The term with the highest marginal utility (Term 3) is identified. Terms 3 and 2 and 1 are the selected query modification terms at this point.
4) Repeat this process until the fifty term limit is reached.

The purpose of this process is to keep the marginal utility of the most recent terms selected from the relevant and non-relevant categories approximately equal throughout the selection procedure. At the query cut-off point, the marginal utility of terms from the relevant category will be approximately equal to the marginal utility of terms from the non-relevant category. This will provide a good balance between maximizing signal and minimizing noise. Obviously, if signal maximization, or noise minimization, become more important for a given study, the differing thresholds for marginal utility for each category can be incorporated into the selection algorithm.

IV-I-ii-b. Maximizing Relevant Records Retrieved

The objective of this option is to maximize the number of relevant records retrieved, subject to the constraints of a signal-to-noise ratio floor, and a ceiling on the number of query terms allowed. A simple heuristic procedure for solving this problem can be demonstrated with the use of Figure A5 and the Excel Solver (a linear/ non-linear optimization package). The Excel solver requires that three parameters be specified in an optimization problem: 1) an objective function to be maximized or minimized; 2) the constraints on the problem; 3) the variables to be changed. On Figure A5, the matrix cell P4-CUMUL RELEV is the entity to be maximized. An additional column would be entered containing the variables to be changed. These variables are the binary coefficients of the phrases Pi. These

coefficients could assume a value of either 0 or 1 (an integer programming solution).

IV-J. Construct Modified Query

The phrases and phrase combinations selected by the above protocols are added to the query, some existing query terms may be removed, and the final query is inserted into the search engine for the next iteration.

IV-K. Query Expansion for Discovery and Innovation

This final pre-Summary/ Conclusions section provides more detail on query expansion for discovery and innovation (Step 2 in Figure 1, in the main text). The objective is to generalize the query terms while maintaining a delicate balance: the generalized terms bear some relation to the initial core literature retrieval terms while the relation is sufficiently indirect for the two literatures to be considered disjoint. Also, the terms are not overly general such that an unwieldy amount of data is retrieved, or the records retrieved are so distant from those of the core literature that impacts will be minimal (Kostoff, 1994a).

The approach consists of examining each core literature query term, and testing different levels of generalizing the term to insure that the above objectives are met. If the records retrieved from the previous iteration are clustered, then terms for the expanded query should be selected such that each main theme from the clustering is adequately represented in the query. Each term being considered for the expanded query should be tested in the source database (e.g., SCI) for retrieval efficiency of relevant records. In some cases, very general forms of the term should be inserted in the source database, and the retrieved records analyzed for more specific variants of the overly general form of the term.

At this point, an example may be illuminating. Consider the topics represented in Figure 1 of the main text. The discovery objective is to expand the core water purification literature to include a more general directly and indirectly-related literature containing potential discovery and innovation candidates. Clustering of the core literature may show very narrow and specific aspects of the assumed two main themes: distillation, membrane filtering ***(mass separation)***, ozonation, chlorination

(disinfection). Phrases selected for the query should be drawn from each of the more generic versions of the main themes.

For example, suppose WATER PURIFICATION is a query term for retrieving the core literature. How could it be generalized for expansion, according to the principles set forth above? One approach is incremental generalization. WATER is a sub-set of LIQUID. Therefore, WATER PURIFICATION could be generalized incrementally to LIQUID PURIFICATION. Use of this query term in the source literature would retrieve documents on purification of liquids in addition to water, and any novel concepts used to purify liquids other than water could be extrapolated to help solve the water purification problem. Before adding this more general term to the query, LIQUID PURIFICATION should be inserted into the SCI search engine, and the retrieval sampled to insure that a high fraction of records relevant to mass separation are being retrieved. In turn, LIQUID is a sub-set of FLUID. Therefore, LIQUID PURIFICATION could be generalized incrementally to FLUID PURIFICATION, and now concepts from the additional gas purification documents could be extrapolated to water purification improvements. The same following steps above would be repeated. The next generalization might be to MASS PURIFICATION, and so on.

How broadly should a query term be generalized? The more directly related the expanded literature is to the core literature, the more obvious will be the connections, but the lower will be the probability for radical discovery and innovation. The more indirectly related the expanded literature is to the core literature, the less obvious will be the connections, but the higher will be the probability for radical discovery and innovation. Thus, if radical discovery and innovation is the goal, the broadest expansion consistent with available resources and reasonable numbers of links in the relational chain should be utilized.

A second approach to generalization for expansion is to filter the core literature records for all phrases containing the word PURIFICATION. Then, each reasonable query expansion term candidate based on PURIFICATION can be checked following the steps above.

A third approach to generalization is to select one of the phrase words that is too generic to use as a stand-alone query term (e.g., WATER or PURIFICATION) for further examination. Since purification is the

technology of interest, the word PURIFICATION would be inserted into the SCI search engine, and thousands of records retrieved. Text analyses would be performed on these retrieved records, and all phrases containing the word PURIFICATION would be extracted, and examined. Each PURIFICATION phrase variant proposed for the query would follow the same checks described above. While time consuming, this is the author's preferred approach for examining foundational terms for query expansion. In a water purification study, for example, this approach could be used for the very generic terms of foundational importance to the core separation processes (e.g., REMOVAL, SEPARATION, PURIFICATION, EXTRACTION, etc). This approach can provide quite comprehensive query terms. Because of the time involved for this latter expansion approach, only the most important generic roots should be examined. All other terms could be examined for expansion using the first or second methods described.

How large should resultant queries be? For queries whose objective is retrieval of a statistically representative sample of documents from the source literature to define the core literature, our marginal utility approach can be used to determine cutoff (Kostoff et al, 2004a). Basically, more query terms provide increased refinement of a fixed scope literature (e.g., the existing water purification literature). In the reference example, a 156 term query was reduced to 100 terms, with essentially no loss in fidelity of retrieval of the NonLinear Dynamics literature. Other queries used in the past for this objective ranged from a handful of terms to hundreds of terms, depending strongly on the technology being examined.

For queries whose objective is retrieval of potential discovery items, most comprehensive retrieval coverage of an expanding scope literature, consistent with high retrieval precision (mainly relevant records), is required. The numbers of query terms will be higher than in the first case, and queries up to many hundreds of terms in length (depending on the specific technologies being studied) are possible, and in fact have been generated.

V. Summary and Conclusions

The origins and recent upgrades of Simulated Nucleation, a Database Tomography-based Relevance Feedback information retrieval process incorporating Term-Co-occurrence and Query Expansion, were described. The new capabilities added to Simulated Nucleation allow the following:

*More relevant records to be obtained
*More non-relevant records to be eliminated
*More rapid and precise identification of desirable query modification terms
*Reduced potential for duplication of query terms
*Selection of query terms with highest marginal utility

New research on Simulated Nucleation should be focused on the following:

*Reducing number of iterative steps
*Reducing sample size per step
*Demonstrating semi-automated optimized query selection technique
*Using additional semantic and syntactic text properties for determining congruency
*Experimenting with gradations of relevance rather than just binary for selecting ‘binning’ categories

Simulated Nucleation is an effective tool for query expansion as a basis for discovery and innovation.

VI. References - Appendix

Attar, R. and Fraenkel, A.S., "Local Feedback in Full-Text Retrieval Systems", Journal of the ACM, 24:3, 1977.

Braun, T., Schubert, A., and Kostoff, R. N. “A Chemistry Field in Search of Applications: Statistical Analysis of U. S. Fullerene Patents”. Journal of Chemical Information and Computer Science. 42:5. 1011-1015. 2002a.

Braun, T., Schubert, A. P., and Kostoff, R. N. "Growth and Trends of Fullerene Research as Reflected in its Journal Literature." Chemical Reviews. 100:1. 23-27. January 2000b.

Callan, J., Croft, W.B., and Broglio, J., "TREC and TIPSTER Experiments with INQUERY", Information Processing and Management, 31:3, 1995.

Chen, H. C., et al. , “A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System”, Journal of the American Society for Information Science. 48:1, 1997.

Chen, H. C., et al. , “Alleviating Search Uncertainty through Concept Associations: Automatic Indexing, Co-occurrence Analysis, and Parallel Computing”, Journal of the American Society for Information Science. 49:3, 1998.

Chung, Y.M., Lee, J.Y. “Optimization of Some Factors Affecting the Performance of Query Expansion”. Information Processing and Management. 40:6. 891-917. Nov 2004.

Croft, W., and Harper, D., "Using Probabilistic Models of Document Retrieval Without Relevance Information", Journal of Documentation, 35, 1979.

Crouch, C. J. , “An Approach to the Automatic Construction of Global Thesauri”, Information Processing and Management, 26:5, 1990.

Del Río, J. A., Kostoff, R. N., García, E. O., Ramírez, A. M., and Humenik, J. A. “Citation Mining Citing Population Profiling using Bibliometrics and Text Mining”. Centro de Investigación en Energía, Universidad Nacional

Autonoma de Mexico. http://www.cie.unam.mx/W_Reportes. 2000.

Del Rio, J. A., Kostoff, R. N., Garcia, E. O., Ramirez, A. M., and Humenik, J. A. "Phenomenological Approach to Profile Impact of Scientific Research: Citation Mining." Advances in Complex Systems. 5:1. 19-42. 2002.

Furnas, G. W., et al, "The Vocabulary Problem in Human- System Communication", Communications of the ACM, 30:11, 1987.

Gomez, L. M., Lochbaum, C. C., and Landauer, T. K. , "All the Right Words: Finding What You Want as a Function of the Indexing Vocabulary, Journal of the American Society for Information Science, 41, 1990.

Harman, D., "Relevance Feedback and Other Query Modification Techniques", in Information Retrieval: Data Structures and Algorithms, Frakes, W. B., and Baeza-Yates, R., Eds., (Prentice-Hall, Englewood Cliffs, NJ), 1992.

Hartley, J. and Kostoff, R. N. "How Useful are 'Key Words' in Scientific Journals?" Journal of Information Science. 29:5. 433-438. October 2003h.

Jing, Y. and Croft, W.B., "An Association Thesaurus for Information Retrieval", in Proceedings of RIAO 94, 1994.

Kostoff, R. N., "Word Frequency Analysis of Text Databases", ONR Memorandum 5000 Ser 10P4/ 1443, April 12, 1991a.

Kostoff, R. N., "Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis," Proceedings: Portland International Conference on Management of Engineering and Technology, October 27-31, 1991b.

Kostoff, R. N., "Co-Word Analysis," in Assessing R&D Impacts: Method and Practice, Bozeman, B. and Melkers, J., Eds. (Kluwer Academic Publishers, Norwell, MA) 1993a.

Kostoff, R. N., "Database Tomography for Technical Intelligence", Proceedings: Eighth Annual Conference of the Society for Competitive Intelligence Professionals, Los Angeles, CA 1993b.

Kostoff, R. N., "Database Tomography for Technical Intelligence," Competitive Intelligence Review, 4:1. 38-43. Spring 1993c.

Kostoff, R.N. and Eberhart, H.J., "Database Tomography: Applications to Technical Intelligence," Proceedings: Technology 2003, Vol. 2, Anaheim, CA, Dec. 7-9, 1993d.

Kostoff, R. N., "Research Impact Quantification," R&D Management, 24:3, July 1994a.

Kostoff, R.N., "Database Tomography: Origins and Applications," Competitive Intelligence Review, Special Issue on Technology, 5:1. 48-55. Spring 1994b.

Kostoff, R. N. and Eberhart, H. J., "Database Tomography: Applications to Information, Logistics, and Personnel Management", Proceedings: Advanced Information Systems & Technology for Acquisition, Logistics, & Personnel Applications, Williamsburg, VA, March 28-30, 1994c.

Kostoff, R. N., "Database Utilization for Research Assessment", The Journal of Information Technology Management, VI:2. 1995a.

Kostoff, R. N., Eberhart, H. J., and Miles, D., "System and Method for Database Tomography", U. S. Patent Number 5440481, August 8, 1995b.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Information Retrieval", Journal of Information Science, 23:4, 1997a.

Kostoff, R. N., "Database Tomography for Technical Intelligence: Analysis of the Research Impact Assessment Literature", Competitive Intelligence Review, 8:2, Summer 1997b.

Kostoff, R. N., Eberhart, H.J., Toothman, D.R., and Pellenbarg, R. "Database Tomography for Technical Intelligence: Comparative Analysis of the Research Impact Assessment Literature and the Journal of the American Chemical Society", Scientometrics, 40:1, 1997c.

Kostoff, R. N., "Citation Analysis Cross-Field Normalization: A New Paradigm", Scientometrics, 39:3, 1997d.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature". Information Processing and Management. 34:1. 69-85. 1998a.

Kostoff, R. N. "The Under-reporting of Research Impact". The Scientist. 14 September 1998b.

Kostoff, R. N. "The Use and Misuse of Citation Analysis in Research Evaluation". Scientometrics. 43:1. September 1998c.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography". Journal of the American Society for Information Science. 50:5. 427-447. 15 April 1999a.

Kostoff, R. N. "Science and Technology Innovation". Technovation. 19:10. 593-604. October 1999b.

Kostoff, R. N., and Geisler, E. "Strategic Management and Implementation of Textual Data Mining in Government Organizations". Technology Analysis and Strategic Management. 11:4. 1999c.

Kostoff, R. N. "Implementation of Textual Data Mining in Government Organizations", Presented at American Society for Information Science Annual Conference. Special Interest Group on Automated Language Processing. 3 November 1999d.

Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. "Fullerene Roadmaps Using Bibliometrics and Database Tomography". Journal of Chemical Information and Computer Science. 40:1. 19-39. Jan-Feb 2000a.

Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". Journal of Aircraft, 37:4. 727-730. July-August 2000c.

Kostoff, R. N. "High Quality Information Retrieval for Improving the Conduct and Management of Research and Development". Proceedings:

Twelfth International Symposium on Methodologies for Intelligent Systems. 11-14 October 2000d.

Kostoff, R. N. "Implementation of Textual Data Mining in Government Organizations". Proceedings: Federal Data Mining Symposium and Exposition. 28-29 March 2000e.

Kostoff, R. N. "The Underpublishing of Science and Technology Results". The Scientist. 14:9. 6-6. 1 May 2000f.

Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. A. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". TR NAWCAD PAX/RTR-2000/84. Naval Air Warfare Center, Aircraft Division, Patuxent River, MD. 2000g.

Kostoff, R. N. "Science and Technology Text Mining". Keynote presentation/ Proceedings. TTCP/ ITWP Workshop. Farnborough, UK. 12 October 2000h.

Kostoff, R. N. "Implementation of Textual Data Mining in Government Organizations". Proceedings: Federal Data Mining Symposium and Exposition, 28-29 March 2000i.

Kostoff, R. N. "The Extraction of Useful Information from the BioMedical Literature". Academic Medicine. 76:12. December 2001a.

Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. "Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling". JASIST. 52:13. 1148-1156. November 2001b.

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., and Humenik, J. A. "Text Mining Using Database Tomography and Bibliometrics: A Review". Technological Forecasting and Social Change. 68:3. November 2001c.

Kostoff, R. N. "Predicting Biowarfare Agents Takes on Priority". The Scientist. 26 November 2001d.

Kostoff, R. N. "Stimulating Discovery". Proceedings: Discovery Science Workshop. November 2001e.

Kostoff, R. N. "Normalization for Citation Analysis". Cortex. 37. 604-606. September 2001f.

Kostoff, R. N., and DeMarco, R. A. "Science and Technology Text Mining". Analytical Chemistry. 73:13. 370-378A. 1 July 2001g.

Kostoff, R. N. "Intel Gold". Military Information Technology. 5:6. July 2001h.

Kostoff, R. N. "Extracting Intel Ore". Military Information Technology. 5:5. 24-26. June 2001i.

Kostoff, R. N., and Schaller, R. R. "Science and Technology Roadmaps". IEEE Transactions on Engineering Management. 48:2. 132-143. May 2001j.

Kostoff, R. N., and Hartley, J. "Structured Abstracts for Technical Journals". Science. 11 May. p.292 (5519):1067a. 2001k.

Kostoff, R. N. "Managing Innovation". Interview. Inside R&D Alert. 12 January 2001l.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography". Journal of Power Sources. 110:1. 163-176. 2002a.

Kostoff, R. N., and Hartley J. "Structured Abstracts for Technical Journals". Journal of Information Science. 28:3. 257-261. 2002b.

Kostoff, R. N. "Citation Analysis for Research Performer Quality". Scientometrics. 53:1. 49-71. 2002c.

Kostoff, R. N. "Biowarfare Agent Prediction". Homeland Defense Journal. 1:4. 1-1. 2002d.

Kostoff, R. N. "Overcoming Specialization." BioScience. 52:10. 937-941. 2002e.

Kostoff, R. N. "Text Mining for Global Technology Watch". In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799. 2003a.

Kostoff, R. N. "Stimulating Innovation". International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 2003b.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. "Fractals Roadmaps using Bibliometrics and Database Tomography". SSC San Diego SDONR 477, Space and Naval Warfare Systems Center. San Diego, CA. June 2003c.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Electrochemical Power: Military Requirements and Literature Structure." Academic and Applied Research in Military Science. 2:1. 5-38. 2003d.

Kostoff, R. N. "Data – A Strategic Resource for National Security". Academic and Applied Research in Military Science. 2:1. 169-172. 2003e.

Kostoff, R. N. "Bilateral Asymmetry Prediction". Medical Hypotheses. 61:2. 265-266. August 2003f.

Kostoff, R.N. "Role of Technical Literature in Science and Technology Development and Exploitation." Journal of Information Science. 29:3. 223-228. 2003g.

Kostoff, R. N. "The Practice and Malpractice of Stemming". JASIST. 54: 10. June 2003i.

Kostoff, R. N., Karpouzian, G., and Malpohl, G. "Abrupt Wing Stall Roadmaps Using Database Tomography and Bibliometrics". TR NAWCAD PAX/RTR-2003/164 Naval Air Warfare Center, Aircraft Division, Patuxent River, MD. 2003j.

Kostoff, R. N. "Science and Technology Text Mining: Cross-Disciplinary Innovation". DTIC Technical Report Number ADA414807, 20 June 2003k.

Kostoff, R. N., and DeMarco, R. A. "Science and Technology Text Mining: Analytical Chemistry". DTIC Technical Report Number ADA415945. 2003l.

Kostoff, R. N. "Science and Technology Text Mining: Management Decision Aids". DTIC Technical Report Number ADA415501. 2003m.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Science and Technology Text Mining: Electrochemical Power." DTIC Technical Report Number ADA415885. 2003n.

Kostoff, R. N. "Science and Technology Text Mining: Global Technology Watch". DTIC Technical Report Number ADA415863. 2003o.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Science and Technology Text Mining: Near-Earth Space". DTIC Technical Report Number ADA415928. 2003p.

Kostoff, R. N., Boylan, R., and Simons, G. R. "Disruptive Technology Roadmaps". DTIC Technical Report Number ADA415933. 2003q.

Kostoff, R. N. "Science and Technology Text Mining: Origins of Database Tomography and Multi-Word Clustering". DTIC Technical Report Number ADA416268. 2003r.

Kostoff, R. N., "Science and Technology Text Mining: Comparative Analysis of the Research Impact Assessment Literature and the Journal of the American Chemical Society." DTIC Technical Report Number ADA416267. 2003s.

Kostoff, R. N., and Hartley, J. "Science and Technology Text Mining: Structured Papers". DTIC Technical Report Number ADA417220 2003t.

Kostoff, R. N., Shlesinger, M., and Tshiteya, R. "Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography". International Journal of Bifurcation and Chaos. 14:1. 61-92. January 2004a.

Kostoff, R. N., Boylan, R., and Simons, G. R. "Disruptive Technology Roadmaps". Technological Forecasting and Social Change. 71:1-2. 141-159. January-February 2004b.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. “Fractals Roadmaps using Bibliometrics and Database Tomography”. Fractals. 12:1. 1-16. March 2004c.

Kostoff, R.N., Bedford, C.W., Del Rio, J. A ., Cortes, H., and Karypis, G. “Macromolecule Mass Spectrometry: Citation Mining of User Documents”. Journal of the American Society for Mass Spectrometry. 15:3. 281-287. March 2004d.

Kostoff, R. N. “Global Technology Watch”. CHIPS Magazine. Summer 2004e.

Kostoff, R. N., Block, J. A., Stump, J. A., and Pfeil, K. M. “Information Content in Medline Record Fields”. International Journal of Medical Informatics. 73:6. 515-527. June. 2004f.

Kostoff, R.N. “Scientific Impact of Nations”. The Scientist. 27 September 2004g.

Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. “Science and Technology Text Mining: Citation Mining of Dynamic Granular Systems.” DTIC Technical Report Number ADA418862. 2004h.

Kostoff, R. N., Bedford, C., Del Rio, J. A., Cortes, H., and Karypis, G. “Macromolecule Mass Spectrometry: Citation Mining of User Documents.” DTIC Technical Report Number ADA418841. 2004i.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Science and Technology Text Mining: Hypersonic and Supersonic Flow". DTIC Technical Report Number ADA418717. 2004j.

Kostoff, R. N., and Geisler, E. “Science and Technology Text Mining : Strategic Management and Implementation in Government Organizations.” DTIC Technical Report Number ADA421060. 2004k.

Kostoff, R. N., Shlesinger, M., and Tshiteya, R. “Science and Technology Text Mining: Nonlinear Dynamics”. DTIC Technical Report Number ADA420998. 2004l.

Kostoff, R. N., Tshiteya, R., Humenik, J. A., and Pfeil, K M. “Science and Technology Text Mining: Electric Power Sources”. DTIC Technical Report Number ADA421789. 2004m.

Kostoff, R. N., Andrews, J., Buchtel, H., Pfeil, K., Tshiteya, R., and Humenik, J. A. “Science and Technology Text Mining: Cortex”. DTIC Technical Report Number ADA425 056. 2004n.

Kostoff, R. N., Block, J. A., Stump, J. A., and Pfeil, K. M. “Information Content in Medline Record Fields”. DTIC Technical Report Number ADA423900. 2004o.

Kostoff, R. N. and Block, J. A. “Context-Dependent Conflation, Text Filtering and Clustering”. DTIC Technical Report Number ADA426072. 1 September 2004p.

Kostoff, R.N. “Science and Technology Citation Analysis: Is Citation Normalization Realistic?” DTIC Technical Report Number ADA426271. 8 September 2004q.

Kostoff, R. N. “A Method for Data and Text Mining and Literature-Based Discovery”. Patent Application Publication Number US 2004/0064438 A1. 1 April 2004r.

Kostoff, R. N., Karpouzian, G., and Malpohl, G. "Abrupt Wing Stall Roadmaps Using Database Tomography and Bibliometrics". Journal of Aircraft. 2005a. In Press.

Kostoff, R. N., Tshiteya, R., Pfeil, K M., Humenik, J. A., and Karypis, G. “Power Source Roadmaps Using Database Tomography and Bibliometrics”. Energy. 30:5. 709-730. 2005b.

Kostoff, R. N., and Block, J. A. “Factor Matrix Text Filtering and Clustering.” JASIST. 2005c. In Press.

Kostoff, R.N., and Shlesinger, M. F. “CAB-Citation-Assisted Background.” Scientometrics. 62:2. 199-212. 2005d.

Kostoff, R. N. “Exploiting Global Science and Technology”. Marine Corps Gazette. 89:3. 56-58. March 2005e.

Kostoff, R. N., Buchtel, H., Andrews, J., and Pfeil, K. "The hidden structure of neuropsychology: Text Mining of the Journal *Cortex*: 1991-2001". Cortex. 41:2. 103-115. April 2005f.

Kostoff, R. N. and Martinez, W.L. "Is Citation Normalization Realistic?" Journal of Information Science. 31:1. 57-61. 2005g.

Kostoff, R. N., Del Rio, J. A., Smith, C., Smith, A., Wagner, C.S., Malpohl, G., Karypis, G., and Tshiteya, R. "Mexico Technology Assessment using Text Mining." Technological Forecasting and Social Change. In Press. 2005h.

Kostoff, R. N. "Science and Technology Knowledge Management". in, New Frontiers on Knowledge Management. (Ed.) Kevin DeSouza. Palgrave Macmillan, United Kingdom. In Press. 2005i.

Kwok, K., "A Network Approach to Probabilistic Information-Retrieval", ACM Transactions on Information Systems, 13:3, 1995.

Lesk, M. (1969) Word-word Associations in Document Retrieval Systems. American Documentation. 20.

Losiewicz, P., Oard, D., and Kostoff, R. N. "Textual Data Mining to Support Science and Technology Management". Journal of Intelligent Information Systems. 15. 99-119. 2000.

Losiewicz, P., Oard, D., and Kostoff, R. N. "Science and Technology Text Mining: Basic Concepts". DTIC Technical Report Number ADA415886. 2003.

Maron, M. and Kuhns, J. "On Relevance, Probabilistic Indexing, and Information Retrieval", Journal of the ACM, 7, 1960.

Rasmussen, E., "Clustering Algorithms", in Frakes, W. B., and Baeza-Yates, R., (Eds), Information Retrieval: Data Structures and Algorithms, Englewood Cliffs, NJ: Prentice Hall, 1992.

Robertson, S. and Sparck Jones, K., "Relevance Weighting of Search Terms", Journal of the American Society for Information Science. 27, 1976.

Rocchio, J., “Relevance Feedback in Information Retrieval. The Smart System-Experiments in Automatic Document Processing”, Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.

Ruthven, I, and Lalmas, M. “A Survey on the Use of Relevance Feedback for Information Access Systems”. Knowledge Engineering Review. 18:2. 95-145. June 2003.

Salton, G., “Relevance Feedback and the Optimization of Retrieval Effectiveness”, The Smart System-Experiments in Automatic Document Processing, Prentice Hall, Inc., Englewood Cliffs, NJ, 1971.

Salton, G., Fox, E., and Vorhees, E., “Advanced Feedback Methods in Information Retrieval”, Journal of the American Society for Information Science, 36, 1985.

Salton, G. and Buckley, C., “Improving Retrieval Performance by Relevance Feedback”, Journal of the American Society for Information Science, 41:4, 1990.

Smeaton, A. and Van Rijsbergen, C., “The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System”, Computer Journal, 26, 1983.

Spink, A., “Term Relevance Feedback and Mediated Database Searching - Implications for Information-Retrieval Practice and Systems-Design”, Information Processing & Management, 31:2, 1995.

Spink, A., Losee, R.M. “Feedback in Information Retrieval”. Annual Review of Information Science and Technology. 31. 33-78. 1996.

Spink, A., Saracevic, T. “Human-Computer Interaction in Information Retrieval: Nature and Manifestations of Feedback”. Interacting With Computers. 10:3. 249-267. June 1998.

Stiles, H., “The Association Factor in Information Retrieval”, Journal of the ACM, 8, 1961.

Van Rijsbergen, C., "Information Retrieval", Butterworths, London, 1979.

Xu, J. and Croft, W.B., “Query Expansion Using Local and Global Document Analysis”, in Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR'96), Zurich, Switzerland, 1996.